Algorithms and hardness results for geometric problems on stochastic datasets

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jie Xue

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Ravi Janardan

July, 2019

Acknowledgements

First of all, my great gratitude goes to my advisor Prof. Ravi Janardan for his continuous support, guidance, and encouragement during my entire Ph.D. life. Without his advice, this thesis can never be completed. I have learned a lot from him about how to do research in theoretical computer science.

Second, I would like to thank Prof. Qie He, Prof. Rui Kuang, Prof. Andrew Odlyzko, and Prof. Shashi Shekhar for serving in my thesis committee and for their valuable feedback to my thesis.

I am also grateful to my lab mates, Akash Agrawal, Rahul Saladi, and Yuan Li, for their friendship and support. By discussing and collaborating with them, I largely deepened my understanding of the research subjects. They helped me a lot during my Ph.D. study in many aspects.

Next, I would like to thank Prof. Haitao Wang, Prof. Timothy M. Chan, and Prof. Pankaj K. Agarwal for hosting my visits to Utah State University, University of Illinois, and Duke University, respectively. These fruitful visits resulted in wonderful collaborations, which broadened significantly the scope of my research.

I would also like to thank the Department of Computer Science & Engineering at the University of Minnesota for the generous funding support over years, including teaching/research assistantships and a Doctoral Dissertation Fellowship.

Furthermore, I am thankful to Prof. Subhash Suri and Prof. Daniel Lokshtanov for offering me a postdoctoral position at University of California, Santa Barbara. I look forward to starting my postdoc life with them after my graduation.

Finally, my deep thanks go to Mr. Fu and Mr. Pan. They are great seers who always guide me to the right direction and give me inspirations. Their wisdom and loftiness have deeply influenced me and benefited me in my life.

Dedication

To my parents, Wenbin Xue and Jiqin Zhou.

Abstract

Traditionally, geometric problems are studied on datasets in which each data object exists with probability 1 at its location in the underlying space. However, in many scenarios, there may be some uncertainty associated with the existence or the locations of the data points. Such uncertain datasets, called *stochastic datasets*, are often more realistic, as they are more expressive and can model the real data more precisely. For this reason, geometric problems on stochastic datasets have received significant attention in recent years. This thesis studies three sets of geometric problems on stochastic datasets equipped with existential uncertainty. The first set of problems addresses the linear separability of a bichromatic stochastic dataset. Specifically, these problems are concerned with how to compute the probability that a realization of a bichromatic stochastic dataset is linearly separable as well as how to compute the expected separation-margin of such a realization. The second set of problems deals with the stochastic convex hull, i.e., the convex hull of a stochastic dataset. This includes computing the expected measures of a stochastic convex hull, such as the expected diameter, width, and combinatorial complexity. The third set of problems considers the dominance relation in a colored stochastic dataset. These problems involve computing the probability that a realization of a colored stochastic dataset does not contain any dominance pair consisting of two different-colored points. New algorithmic and hardness results are provided for the three sets of problems.

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Figures vii

1 Introduction 1

1.1 Problem statement 2

1.2 Related work 5

1.3 Our contributions 7

1.4 Organization 9

2 Stochastic separability problems 10

2.1 Preliminaries 10

2.2 Separable-probability 12

2.2.1 Extreme separator 12

2.2.2 Computing the separable-probability 16

2.2.3 Improving the SP algorithm 18

2.2.4 Witness-based lower bound for separable-probability 21

2.3 Expected separation-margin 26

2.3.1 Computing the expected separation-margin 28

2.3.2 Improving the ESM algorithm 30

2.3.3 Hardness of computing expected separation-margin 33

2.4 Extension to general geometric objects 36

2.4.1 Reducing polytopes to points 36
2.4.2 Handling balls 37

3 Stochastic convex hull problems **43**
3.1 Preliminaries 43
3.2 Approximating the expected diameter 44
3.2.1 The witness sequence 44
3.2.2 An (n, d)-polynomial-time approximation algorithm 48
3.2.3 A polynomial-time approximation scheme 51
3.2.4 #P-hardness of the expected-diameter problem 53
3.3 Approximating the expected width 55
3.3.1 The witness simplex 55
3.3.2 An $O(1)$-approximation algorithm 58
3.3.3 A fully polynomial-time randomized approximation scheme . 61
3.3.4 A polynomial-time approximation scheme 63
3.4 Computing the expected combinatorial complexity 65
3.4.1 Reduction to SCH membership probability queries 66
3.4.2 Handling the cases $k = d - 2$ and $k = d - 1$ 68
3.4.3 Computing the $\mathcal{S}$-statistics for $\mathcal{E}$ 70

4 Stochastic dominance problems **72**
4.1 Preliminaries 72
4.2 The colored stochastic dominance problem 73
4.2.1 An algorithm for $d = 2$ 73
4.2.2 Hardness results in higher dimensions 82
4.2.3 A simple FPRAS 101
4.3 The free-basis colored stochastic dominance problem 103
4.3.1 Reduction from the CSD problem 103
4.3.2 Reduction to the CSD problem for $d = 2$ 111

5 Conclusion and future work **116**
5.1 Conclusion 116
5.2 Future work 117

References **119**

List of Figures

2.1 Illustrating U^* in $\mathbb{R}^2$. Note that P_1 is not shown to avoid confusion. 13
2.2 Illustrating the extreme separator in $\mathbb{R}^2$. 16
2.3 Illustrating the location of o. The space in the figure is the 2-dim subspace of $\mathbb{R}^d$ that is parallel to the x_1x_2-plane and contains $\hat{r}$, $\hat{b}$. . 18
2.4 Illustrating how to use duality and topological sweep to eliminate the log factor in runtime. 19
2.5 An example of support set and support plane, in $\mathbb{R}^2$ 28
2.6 A separability problem for a set of bichromatic general objects in $\mathbb{R}^2$ 36
3.1 The locations of x, p, q and y in the proof of Lemma 17. 45
3.2 An illustration of $B_u, B_v, B_{v'}$ in the proof of Lemma 18. 47
3.3 The regular double-simplex in the proof of Lemma 24. 54
4.1 Illustrating A and $Z(A)$. 73
4.2 Illustrating $(i,j)_{\searrow}$ and $(i,j)_{\nwarrow}$ for a legal pair (i,j). 78
4.3 Illustrating Lemma 41. The orange color is only used to highlight each range and does *not* represent the color of each point. Dashed (resp. solid) boundaries are exclusive (resp. inclusive). 80
4.4 Inserting new vertices into each edge of G. 87
4.5 An orthogonal grid drawing. 89
4.6 The construction of P_e. 91
4.7 The minimal standard triangle in $\mathbb{R}^2$ containing a set of points. . . . 92
4.8 The case that s is horizontal. 92
4.9 The case that s is vertical. 92
4.10 A local structure of f_p in the box B_i. 97
4.11 Deleting a vertex and adding three new edges. 100

4.12 An example of witness pair. $l = \overline{a_1 - a_8} = \overline{a_2 - a_5}$. $\text{wit}(R) = (a_1, a_8)$. 113

Chapter 1

Introduction

Computational geometry is concerned with design and analysis of efficient algorithms and data structures to model and manipulate geometric objects (such as points, lines, segments, polygons, polyhedra, etc.), which are ubiquitous in the real world [1, 2]. In traditional computational geometry, the geometric data involved is usually assumed to be deterministic; that is, each point (or more generally, geometric object) in the given dataset exists with probability 1 at its location in the underlying space. However, such an assumption is not reasonable in many scenarios. For instance, due to limitations of sensing devices, sometimes the existence or the locations of the data points cannot be determined precisely. Due to this uncertainty, the conventional deterministic dataset may fail to model the real data accurately.

For the purpose of resolving this issue, researchers have generalized the conventional dataset to the so-called *stochastic* dataset (or probabilistic dataset), which allows the data points to have some uncertainty. There are two main models of uncertainty in the literature. In the *existential uncertainty* model, a data point has a known location but can have an uncertain existence, which is modeled by an *existence probability*. The existential uncertainty model is useful in the case where the data points obtained are not totally reliable (the existence probabilities express the reliabilities of the data points). In the *locational uncertainty* model, a data point can have an uncertain location in the underlying space, which is modeled by a probability distribution (either discrete or continuous). The locational uncertainty model is useful in the case where the locations of the data points cannot be uniquely

determined. Compared with conventional datasets, stochastic datasets can model the real data more precisely, and hence are more preferable in many scenarios. In recent years, stochastic datasets have received considerable attention in computational geometry. Many classical geometric problems have been investigated on stochastic datasets. However, due to the uncertainty, the problems on stochastic datasets are usually quite challenging. For example, the closest-pair problem, which aims to find the pair of points with minimum distance among a given set of n (non-stochastic) data points, can be solved in $O(n \log n)$ time, while many stochastic versions of the closest-pair problem have been proved to be NP-hard or #P-hard even in $\mathbb{R}^2$.

In this thesis, we study several classical geometric problems on stochastic datasets. The first set of problems considers the linear separability of a bichromatic stochastic dataset. The second set of problems considers the convex hull of a stochastic dataset, which we call stochastic convex hull (SCH). The third set of problems considers the dominance relation among the points in a stochastic dataset. These problems are described in detail in Section 1.1. We investigate these problems on stochastic datasets equipped with existential uncertainty, and present new algorithmic and hardness results. We remark that some of our algorithms can be extended to more general uncertainty models (e.g., the discrete distribution model in which the existence and location of each data point are both uncertain and the probabilistic location of each point is depicted by a discrete distribution).

1.1 Problem statement

In order to describe the problems we study, we need to formally define stochastic datasets (equipped with existential uncertainty) and some related notions. A *stochastic dataset* in $\mathbb{R}^d$ is a pair $\mathcal{S} = (S, \pi)$ where $S \subseteq \mathbb{R}^d$ is a finite set of points and $\pi : S \to (0, 1]$ is a function indicating the *existence probability* of each point in S. A realization of $\mathcal{S}$ is a random subset of S obtained by including each point $a \in S$ independently with its existence probability $\pi(a)$. A *bichromatic dataset* in $\mathbb{R}^d$ is a pair $\mathcal{T} = (T_R, T_B)$ where $T_R \subseteq \mathbb{R}^d$ (resp. $T_B \subseteq \mathbb{R}^d$) is a finite set of red (resp., blue) points. The *size* of $\mathcal{T}$ is a pair (n, N) of integers where $n = \min\{|T_R|, |T_B|\}$ and $N = \max\{|T_R|, |T_B|\}$. A *subset* of $\mathcal{T}$ is a bichromatic dataset $\mathcal{T}' = (T'_R, T'_B)$ where $T'_R \subseteq T_R$ and $T'_B \subseteq T_B$. A *bichromatic stochastic dataset* in $\mathbb{R}^d$ is a triple

$\mathcal{S} = (S_R, S_B, \pi)$ where (S_R, S_B) is a bichromatic dataset and $\pi : S_R \cup S_B \to (0, 1]$ is the existence-probability function. The *size* of $\mathcal{S}$ is the size of the bichromatic dataset (S_R, S_B). A *realization* of $\mathcal{S}$ is a random subset of (S_R, S_B) obtained by including each point $a \in S_R \cup S_B$ independently with its existence probability $\pi(a)$. By further generalizing the notion of bichromatic datasets, we can define colored datasets and colored stochastic datasets. A *colored dataset* in $\mathbb{R}^d$ is a pair $\mathcal{T} = (T, \mathrm{cl})$ where $T \subseteq \mathbb{R}^d$ is a finite set of points and $\mathrm{cl} : T \to \mathbb{N}$ is the *coloring* (or *coloring function*) indicating the color labels of the points. A *subset* of $\mathcal{T}$ is a colored dataset $\mathcal{T}' = (T', \mathrm{cl}')$ where $T' \subseteq T$ and $\mathrm{cl}' = \mathrm{cl}_{|T'}$, i.e., cl restricted to T'. A *colored stochastic dataset* in $\mathbb{R}^d$ is a triple $\mathcal{S} = (S, \mathrm{cl}, \pi)$ where (S, cl) is a colored dataset and $\pi : S \to (0, 1]$ is the existence-probability function. A *realization* of $\mathcal{S}$ is a random subset of (S, cl) obtained by including each point $a \in S$ independently with its existence probability $\pi(a)$.

Stochastic separability. Linear separability, which is concerned with whether a bichromatic dataset can be separated by a hyperplane into two sets (one of each color), is a basic notion studied in computational geometry, and has many applications. It is also strongly related to the classification task in machine learning and data mining. In this thesis, we study two problems regarding the linear separability of a given bichromatic stochastic dataset $\mathcal{S} = (S_R, S_B, \pi)$ in $\mathbb{R}^d$, both of which are natural generalizations of the classical linear separability problems.

- **Separable-probability.** The first problem aims to compute the *separable-probability* (SP) of a realization of $\mathcal{S}$, i.e., the probability that a realization of $\mathcal{S}$ is linearly separable by a hyperplane. A bichromatic dataset is linearly separable if there exists a hyperplane h such that the red points and blue points are on opposite sides of h.

- **Expected separation-margin.** The second problem aims to compute the *expected separation-margin* (ESM) of a realization of $\mathcal{S}$. Roughly speaking, the separation-margin of a (separable) bichromatic dataset is the maximum distance between a separator and the data points. (This notion will be formally defined in Chapter 3.)

Stochastic convex hull. The *convex hull* of a set A of points is, by definition, the smallest convex set containing A [3]. It is one of the most fundamental structures

in computational geometry and has a wide range of applications in areas as diverse as computer graphics, pattern recognition, statistics, robotics, and computer-aided design, among others. A *stochastic convex hull* (SCH) refers to the convex hull of a realization of a stochastic dataset, which is a probabilistic convex polytope. In this thesis, we study three problems regarding a SCH of a given stochastic dataset $\mathcal{S} = (S, \pi)$ in $\mathbb{R}^d$, each of which aims to compute the expectation of some basic statistic of a SCH.

- **Expected diameter.** The first problem aims to (approximately) compute the expected diameter of a SCH of $\mathcal{S}$. The *diameter* of a convex polytope is the maximum distance between its two vertices.
- **Expected width.** The second problem aims to (approximately) compute the expected width of a SCH of $\mathcal{S}$. The *width* of a convex polytope is the minimum distance between two parallel hyperplanes that enclose it.
- **Expected combinatorial complexity.** The third problem aims to compute the expected combinatorial complexity of a SCH of $\mathcal{S}$. The *combinatorial complexity* of a convex polytope is the total number of its faces (of dimensions $0, 1, \ldots, d-1$).

Stochastic dominance. A point $p \in \mathbb{R}^d$ is said to *dominate* another point $q \in \mathbb{R}^d$ if the coordinate of p is greater than or equal to the coordinate of q in every dimension. The dominance relation is an important notion in multi-criteria decision-making, and has been well-studied in computational geometry, database, optimization, and other related areas. In this we study two problems regarding the dominance relation among the points in a realization of a given colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ in $\mathbb{R}^d$.

- **Inter-color dominance-free probability.** The first problem aims to compute the probability that a realization of $\mathcal{S}$ does not contain an *inter-color dominance pair*, that is, a pair of points of distinct colors in which one point dominates the other. We call this the *colored stochastic dominance* (CSD) problem.
- **Free-basis inter-color dominance-free probability.** The second problem aims to compute the probability that a realization of $\mathcal{S}$ does not contain an

inter-color dominance pair with respect to some orthogonal basis of $\mathbb{R}^d$. We call this the *free-basis colored stochastic dominance* (FBCSD) problem.

1.2 Related work

The study of geometric problems under uncertainty is a relatively new topic, and has attracted a lot of attention in recent years. Many classical geometric problems have been investigated on stochastic datasets, e.g., nearest-neighbor search [4, 5, 6], convex hulls [7, 8, 9, 10, 11, 12], minimum spanning trees [13], closest pair [8, 14, 6], range search [15, 16], clustering [17], Voronoi diagrams [18, 5], line arrangements [19], line separability [20, 21, 22], skylines [23], dominance relations [24], etc. In what follows, we sample some existing results which are strongly relevant to this thesis.

Stochastic separability. Linear separability related-problems have been well studied for years in computational geometry, and also arisen during data classification in machine learning and data mining. Linear separability on stochastic datasets has been studied in [20, 21, 22]; this thesis presents the results in [22]. The work [20] considered some separability problems in $\mathbb{R}^2$ where the locations of the points are uncertain: it is assumed that each point is drawn uniformly from an axis-parallel rectangle. Specifically, the work [20] studied how to find certain separators, possible separators, most-likely separators, and maximal separators, for such a set of uncertain points. In [21], Fink et al. studied the separable-probability problem in $\mathbb{R}^d$ under existential uncertainty, which is also investigated in this thesis (and thus in [22]). An $O(nN^{d-1})$-time algorithm was given in [21] to compute the separable-probability of a bichromatic stochastic dataset in $\mathbb{R}^d$ of size (n, N). Our algorithm presented in this thesis and in [22] achieves the same bound, and the two results were in fact obtained simultaneously and independently. In terms of techniques, however, our algorithm is quite different from the algorithm in [21]. The latter algorithm computes the separable-probability by adding a dummy anchor point and using an inclusion-exclusion strategy. On the other hand, our algorithm solves the problem more directly: it does not introduce any additional points and the separable-probability is computed using a simple addition principle. The paper [21]

also gave some hardness results for the separable-probability problem, and a reduction from the SCH membership probability problem to the separable-probability problem.

Stochastic convex hull. The convex hull is one of the most fundamental structures in computational geometry, and has been well-studied over years (see, for example, [3] for a survey). Convex hulls under uncertainty have been studied in [7, 8, 9, 10, 11, 12]; this thesis presents the results in [12] (and some additional results). The work [7] studied how to compute the probability that a given point is inside a SCH, called *SCH membership probability*. The paper [11] considered the problem of finding the most likely convex hull of a stochastic dataset. In [10], Löffler and van Kreveld investigated the largest and smallest convex hull of a set of uncertain points in $\mathbb{R}^2$. More relevantly, the problem of computing the expected diameter of a SCH was studied in [8] and [9]. Huang and Li [8] provided a fully polynomial-time randomized approximation scheme (FPRAS) for computing the expected farthest-pair distance of a stochastic dataset in a metric space, which directly implies an FPRAS for computing the expected diameter of a SCH, since in Euclidean space the farthest-pair distance of a set of points is just the diameter of their convex hull. Li et al. [9] gave a deterministic $(2/\sqrt{3})$-approximation algorithm for computing the expected diameter of a SCH, which is based on an (exact) algorithm for computing the expected diameter of the stochastic smallest enclosing ball. Although the work [9] only considered the case in $\mathbb{R}^2$, the algorithm can be naturally extended to compute a $(\sqrt{2d}/\sqrt{d+1})$-approximation of the expected diameter of a SCH in $\mathbb{R}^d$. Nevertheless, the runtime of this algorithm grows exponentially as d increases, since computing the expected diameter of the stochastic smallest enclosing ball requires $n^{\Omega(d)}$ time [25]. The width and combinatorial complexity of a SCH had not yet been investigated previously, to the best of our knowledge.

Stochastic dominance. Classical studies regarding the dominance relation can be found in many works such as [26, 27, 28]. Recently, there have been efforts to consider the dominance relation on stochastic datasets [29, 23, 30, 31]. The main focus of these efforts is the behavior of the skyline points (i.e., the points that are not dominated by any other points) of a stochastic dataset. The problems and results presented in this thesis are based on the work reported in the manuscript

[24]. To the best of our knowledge, these problems, which consider the probability that a realization of a stochastic probability is dominance-free, have not been studied before.

1.3 Our contributions

In this thesis, we present new algorithms for the three sets of problems defined in Section 1.1, as well as some hardness results. In most of the problems studied, we assume the dimension d is a fixed constant.

Stochastic separability. We study the separable-probability (SP) problem and the expected separation-margin (ESM) problem, which are defined in Section 1.1. We obtain the following results.

- We give an $O(nN^{d-1})$-time (resp., $O(\min\{nN \log N, N^2\})$-time) algorithm for computing the SP of a given bichromatic stochastic dataset in $\mathbb{R}^d$ of size (n, N) for $d \geq 3$ (resp., $d = 2$). An application of this algorithm to the SCH membership probability problem is provided. On the other hand, we show that the time complexity of any so-called witness-based algorithm for the SP problem in $\mathbb{R}^d$ is $\Omega(nN^{d-1})$ for $d \geq 3$.

- We propose an $O(nN^d)$-time algorithm for computing the ESM of a given bichromatic stochastic dataset in $\mathbb{R}^d$ of size (n, N) for $d \geq 2$. We also provide a hardness result showing that further improving our algorithm might be difficult.

- We show that our algorithms above can be extended to solve the separability problems for bichromatic stochastic datasets consisting of general geometric objects (specifically, polytopes of constant complexity and balls), resulting in an $O(nN^d)$-time SP algorithm and an $O(nN^{d+1})$-time ESM algorithm.

Stochastic convex hull. We study how to compute the expected diameter, width, and combinatorial complexity of a SCH, which are defined in Section 1.1. We obtain the following results.

- We give a 1.633-approximation algorithm for computing the expected diameter of a SCH of a given stochastic dataset in $\mathbb{R}^d$ of size n. The time complexity

of the algorithm is (n, d)-polynomial (i.e., polynomial in both n and d); here we do *not* assume d is a fixed constant. We also provide a polynomial-time approximation scheme (PTAS) for computing the expected diameter when d is a constant. Finally, we prove that, when d is not a constant, computing the expected diameter exactly is #P-hard. Roughly speaking, the complexity class #P consists of the problems that can be formulated as counting the number of accepting paths of a polynomial-time non-deterministic Turing machine. A problem is #P-*hard* if every other problem in #P can be reduced to it in polynomial time.

- We propose an $O(n^{d+1} \log n)$-time constant-approximation algorithm for computing the expected width of a SCH of a given stochastic dataset in $\mathbb{R}^d$ of size n, when d is a constant. We also provide a fully polynomial-time randomized approximation scheme (FPRAS) and a PTAS for the expected width when d is a constant.

- We give an $O(n^d)$-time exact algorithm for computing the expected combinatorial complexity of a SCH of a given stochastic dataset $\mathbb{R}^d$ of size n, when d is a constant.

Stochastic dominance. We study the colored stochastic dominance (CSD) problem and the free-basis colored stochastic dominance (FBCSD) problem in $\mathbb{R}^d$, as defined in Section 1.1. We obtain the following results.

- We give an $O(n^2 \log^2 n)$-time exact algorithm to solve the CSD problem for $d = 2$. On the other hand, we show that even the CSD problem with a restricted color pattern is #P-hard for $d \geq 3$. Furthermore, even if the existence probabilities of the points are restricted to be $\frac{1}{2}$, the problem remains #P-hard for $d \geq 7$. We also give a FPRAS for the problem in any dimension.

- We show that the CSD problem is polynomial-time reducible to the FBCSD problem in the same dimension, which implies the #P-hardness of the latter for $d \geq 3$. For $d = 2$, we give an $O(n^4 \log^2 n)$-time exact algorithm for solving the FBCSD problem.

1.4 Organization

The rest of the thesis is organized as follows. Chapter 2 presents our results for the stochastic separability problems. Chapter 3 presents our results for the stochastic convex hull problems. Chapter 4 presents our results for the stochastic dominance problems. Finally, in Chapter 5, we conclude the thesis and suggest some potential directions for future study.

Chapter 2

Stochastic separability problems

Let $\mathcal{S} = (S_R, S_B, \pi)$ be a given bichromatic stochastic dataset in $\mathbb{R}^d$, and (n, N) be the size of $\mathcal{S}$. Set $S = S_R \cup S_B$. In this chapter, we study the problems of computing the separable-probability and the expected separation-margin of $\mathcal{S}$; see Section 1.1 for the statement of these problems.

2.1 Preliminaries

Let $\mathcal{T} = (T_R, T_B)$ be a bichromatic dataset in $\mathbb{R}^d$. We say $\mathcal{T}$ is *strongly separable* if there exists a hyperplane h such that all the points in T_R are in one connected component of $\mathbb{R}^d \backslash h$ while all the points in T_B are in the other connected component of $\mathbb{R}^d \backslash h$; we call h a *strong separator* of $\mathcal{T}$. Also, We say $\mathcal{T}$ is *weakly separable* if there exists a hyperplane h such that, except the points lying on h, all the points in T_R are in one connected component of $\mathbb{R}^d \backslash h$ while all the points in T_B are in the other connected component of $\mathbb{R}^d \backslash h$; we call h a *weak separator* of $\mathcal{T}$. Note that when the points in $T_R \cup T_B$ are in general position, $\mathcal{T}$ is strongly separable iff $\mathcal{T}$ is weakly separable. The following classical lemma gives a criterion for the strong separability of a bichromatic dataset.

Lemma 1. *A bichromatic dataset $\mathcal{T} = (T_R, T_B)$ is strongly separable iff $\mathcal{CH}(T_R) \cap \mathcal{CH}(T_B) = \emptyset$, where $\mathcal{CH}(\cdot)$ denotes the convex hull.*

Proof. We first prove the "only if" part. Suppose we have a strong separator h for $\mathcal{T}$. Let H be the half space bounded by h which contains T_R and H' be the

other half space bounded by h which contains T_B. Since both $H\backslash h$ and $H'\backslash h$ are convex, we have $\mathcal{CH}(T_R) \subseteq H\backslash h$ and $\mathcal{CH}(T_B) \subseteq H'\backslash h$. It immediately follows that $\mathcal{CH}(T_R) \cap \mathcal{CH}(T_B) = \emptyset$. To prove the "if" part, assume $\mathcal{CH}(T_R) \cap \mathcal{CH}(T_B) = \emptyset$. Let (r, b) be the closest pair of points where $r \in \mathcal{CH}(T_R)$ and $b \in \mathcal{CH}(T_B)$. We denote the midpoint of the segment $[r, b]$ by s and define h as the hyperplane going through s and perpendicular to $[r, b]$. We claim that h is a strong separator for $\mathcal{T}$. Assume h does not strongly separate $\mathcal{T}$. That means there are two points in T_R (or T_B) that are not on the same (open) side of h. Without loss of generality, we just assume such two points are in T_R. Thus, we can find a point $r^* \in \mathcal{CH}(T_R)$ that is on h. Consider the triangle $\triangle brr^*$. Since r^* is on h, we have $\angle brr^* < \pi/2$. Therefore, there exists a point t on the segment $[r, r^*]$ such that $\text{dist}(t, b) < \text{dist}(r, b)$. This contradicts the fact that (r, b) is the closest pair, because $t \in \mathcal{CH}(T_R)$. Thus, h is a strong separator for $\mathcal{T}$. □

If h is a separator (either strong or weak) of $\mathcal{T}$, the *margin* $M_h(\mathcal{T})$ of h is defined as $M_h(\mathcal{T}) = \min_{a \in T} \text{dist}(a, h)$, where $T = T_R \cup T_B$. The *separation-margin* $Mar(\mathcal{T})$ of $\mathcal{T}$ is defined as $Mar(\mathcal{T}) = \sup_h M_h(\mathcal{T})$ where h is taken over all separators of $\mathcal{T}$ (if $\mathcal{T}$ is trivial or is not separable, we set $Mar(\mathcal{T}) = 0$). We say a separator h of $\mathcal{T}$ is a *maximum-margin separator* if its margin is equal to $Mar(\mathcal{T})$.

If U is a $(d-1)$-dim linear subspace (i.e., a hyperplane) of $\mathbb{R}^d$ and $X \subseteq \mathbb{R}^d$ is a set of points, we write $X^U = \{p(x) : x \in X\}$, where $p : \mathbb{R}^d \to U$ is the orthogonal projection function; in other words, X^U is the set of points in U obtained by orthogonally projecting the points in X to U. Now we introduce a notion called *derived separator*.

Definition 2. *Let $\mathcal{T} = (T_R, T_B)$ be a bichromatic dataset in $\mathbb{R}^d$, and U be a $(d-1)$-dim linear subspace of $\mathbb{R}^d$. Suppose h is a strong (resp., weak) separator of (T_R^U, T_B^U) in the space U. It is easy to see that the pre-image h' of h under the orthogonal projection function $p : \mathbb{R}^d \to U$ is a strong (resp., weak) separator of $\mathcal{T}$ in $\mathbb{R}^d$. We call h' the* ***derived separator*** *of h in $\mathbb{R}^d$.*

2.2 Separable-probability

We study the problem of computing the separable-probability (SP) of $\mathcal{S}$ (denoted by $SP(\mathcal{S})$), i.e., the probability that a realization of $\mathcal{S}$ is (strongly) separable. Trivially, $SP(\mathcal{S})$ can be computed by simply enumerating all the 2^{n+N} possible realizations of $\mathcal{S}$ and summing up the probabilities of the separable ones, which takes exponential time. In order to solve the problem more efficiently than by brute-force, one has to categorize all the separable realizations of $\mathcal{S}$ into a reasonable number of groups such that the sum of the probabilities of the realizations in each group can be easily computed. A natural approach is to charge each separable realization to a unique separator, and use that as the key to do the grouping. The uniqueness requirement here is to avoid over-counting. In addition, all these separators should be easy to enumerate and the sum of the probabilities of those separable realizations charged to each separator should be efficiently computable. In $\mathbb{R}^1$ and $\mathbb{R}^2$, this is easy to achieve. For example, in $\mathbb{R}^1$, given a separable bichromatic dataset, all the possible separators form a segment, and we can choose the leftmost endpoint as the unique separator; in $\mathbb{R}^2$, we can choose the most counterclockwise separator, which goes through exactly one red and one blue point, as the unique separator. It is easy to see that, with the separators chosen above, $SP(\mathcal{S})$ can be easily computed by considering the sum of the probabilities of the realizations charged to each such separator. However, to define such a separator in higher dimensions turns out to be challenging. To solve this problem, we define an important notion called *extreme separator*.

2.2.1 Extreme separator

For convenience, we assume that the points in S have the strong general position property (SGPP), which is defined as follows. Let $I = \{i_1, \ldots, i_{|I|}\}$ be any subset of the index set $\{1, \ldots, d\}$ where $i_1 < \cdots < i_{|I|}$. We define a projection function $\phi_I : \mathbb{R}^d \to \mathbb{R}^{|I|}$ as

$$(x_1, \ldots, x_d) \mapsto (x_{i_1}, \ldots, x_{i_{|I|}}).$$

Also, for any $X \subseteq \mathbb{R}^d$, we define $\Phi_I(X) = \{\phi_I(x) : x \in X\}$. Let A be a set of points in $\mathbb{R}^d$. When $d \leq 2$, we say A has *SGPP* if it is in general (linear) position,

i.e., affinely independent. When $d \geq 3$, we say A has *SGPP* if **(1)** A is in general (linear) position and **(2)** $\Phi_J(A)$ has SGPP for $J = \{3, \ldots, d\}$.

Recall that in $\mathbb{R}^2$ we define the extreme separator of a separable bichromatic dataset as the (weak) separator with the most counterclock-wise counterclockwise position. In the general case, we essentially follow this basic idea: we try to define the extreme separator as a separator with an "extreme" location. However, to find such a separator requires nontrivial effort.

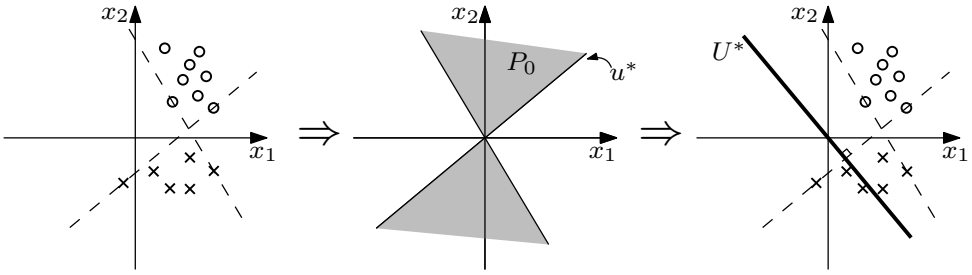


Figure 2.1: Illustrating U^* in $\mathbb{R}^2$. Note that P_1 is not shown to avoid confusion.

Suppose we are given a separable bichromatic dataset $\mathcal{T} = (T_R, T_B)$ in $\mathbb{R}^d$ for $d \geq 2$ such that $T = T_R \cup T_B$ has SGPP. Let $\mathcal{V}$ be the collection of the $(d-1)$-dim linear subspaces of $\mathbb{R}^d$ whose equations are of the form $ax_1 + bx_2 = 0$, where a and b are constants not equal to 0 simultaneously. In other words, $\mathcal{V}$ contains all the $(d-1)$-dim linear subspaces that are perpendicular to the x_1x_2-plane and go through the origin. Then there is a natural one-to-one correspondence between $\mathcal{V}$ and $\mathbb{P}^1$ (i.e., the 1-dim projective space) given by

$$\sigma : [ax_1 + bx_2 = 0] \longleftrightarrow [a : b].$$

For convenience, we use σ to denote the maps in both directions. Define a map $\rho_{\mathcal{T}} : \mathcal{V} \to \{0, 1\}$ as

$$\rho_{\mathcal{T}}(V) = \begin{cases} 1 & \text{if } (T_R^V, T_B^V) \text{ is strongly separable,} \\ 0 & \text{otherwise.} \end{cases}$$

The map $\rho_{\mathcal{T}}$ induces another map $\rho_{\mathcal{T}}^* : \mathbb{P}^1 \to \{0, 1\}$ via the composition $\rho_{\mathcal{T}}^* = \rho_{\mathcal{T}} \circ \sigma$. Let P_0 and P_1 be the pre-images of $\{0\}$ and $\{1\}$ under $\rho_{\mathcal{T}}^*$, respectively (see Figure 2.1). By applying Lemma 1, it is easy to prove the following.

Lemma 3. *P_0 is a connected closed subspace of $\mathbb{P}^1$. Also, $P_0 = \emptyset$ iff $(\Phi_J(T_R), \Phi_J(T_B))$ is strongly separable in $\mathbb{R}^{d-2}$ for $J = \{3, \ldots, d\}$.*

Proof. We define a subset of $\mathcal{CH}(T_R) \times \mathcal{CH}(T_B)$ as

$$D = \{(r, b) \in \mathcal{CH}(T_R) \times \mathcal{CH}(T_B) : \phi_J(r) = \phi_J = (b)\},$$

where $J = \{3, \ldots, d\}$. Also, define a continuous function $f : D \to \mathbb{P}^1$ as

$$f : (r, b) \longmapsto [(r^{(1)} - b^{(1)}) : (r^{(2)} - b^{(2)})],$$

where $r^{(i)}$ and $b^{(i)}$ denote the i-th coordinates of r and b, respectively. We shall first prove that P_0 is equal to the image $\mathrm{Im}f$ of f. Let $u = [a : b]$ be a point in $\mathbb{P}^1$ and $U = \sigma(u)$. According to Lemma 1, $u \in P_0$ iff $\mathcal{CH}(T_R^U) \cap \mathcal{CH}(T_B^U) \neq \emptyset$. It is clear that $\mathcal{CH}(T_R^U) \cap \mathcal{CH}(T_B^U) \neq \emptyset$ iff u is in the image of f, which implies $P_0 = \mathrm{Im}f$. Then it suffices to prove the lemma regarding $\mathrm{Im}f$ instead of P_0. Because of the connectedness and compactness of D, $\mathrm{Im}f$ is also connected and compact. Furthermore, since $\mathbb{P}^1$ is Hausdorff (i.e., any two distinct points in $\mathbb{P}^1$ have disjoint open neighborhoods), $\mathrm{Im}f$ is closed in $\mathbb{P}^1$. Thus, the first statement of the lemma is proved. To prove the second statement, we first assume $\mathrm{Im}f = \emptyset$, which implies $D = \emptyset$. It then immediately follows that $(\Phi_J(T_R), \Phi_J(T_B))$ is strongly separable in $\mathbb{R}^{d-2}$ for $J = \{3, \ldots, d\}$. On the other hand, if $(\Phi_J(T_R), \Phi_J(T_B))$ is strongly separable, $\mathcal{CH}(T_R^U) \cap \mathcal{CH}(T_B^U) = \emptyset$. In this situation, D has to be empty and thus $\mathrm{Im}f = \emptyset$. □

If $P_0 = \emptyset$, we say the extreme separator of $\mathcal{T}$ is *not defined*. Assume $P_0 \neq \emptyset$. Since P_0 is a connected closed subspace of $\mathbb{P}^1$, it has a unique clockwise boundary point u^* (i.e., the last point of P_0 in the clockwise direction). Let $U^* = \sigma(u^*)$ be the linear subspace in $\mathcal{V}$ corresponding to u^* (see Figure 2.1 again). The following lemma reveals the separability property of $\mathcal{T}^{U^*} = (T_R^{U^*}, T_B^{U^*})$.

Lemma 4. *There exists a unique weak separator for $\mathcal{T}^{U^*}$ in U^*. This separator goes through exactly d points in $\mathcal{T}^{U^*}$, of which at least one is in $T_R^{U^*}$ and one is in $T_B^{U^*}$.*

Proof. Suppose that $\alpha_{\vec{v}} = \min_{r \in T_R}\{\vec{v} \cdot r\}$, $\alpha'_{\vec{v}} = \max_{r \in T_R}\{\vec{v} \cdot r\}$, $\beta_{\vec{v}} = \min_{b \in T_B}\{\vec{v} \cdot b\}$, $\beta'_{\vec{v}} = \max_{b \in T_B}\{\vec{v} \cdot b\}$. Define a function $f : \mathbb{P}^1 \to \mathbb{R}$ as

$$f(u) = \sup_{\vec{v} \in \bar{U}} \max\{(\alpha_{\vec{v}} - \beta'_{\vec{v}}), (\beta_{\vec{v}} - \alpha'_{\vec{v}})\},$$

where $\bar{U} = \mathbb{S}^{d-1} \cap \sigma(u)$ ($\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$). It is easy to see that f is continuous. Furthermore, according to the definition of P_0, we know that $u \in P_0$ iff $f(u) \leq 0$. Since u^* is a boundary point of P_0, we have $f(u^*) = 0$. Thus, $\mathcal{T}^{U^*}$ is weakly (but not strongly) separable. To prove there exists a unique weak separator satisfying the desired properties, we introduce a definition called *degree*. Let X be a polytope and x be a point on the boundary of X. We define the degree of x in X, denoted by $\deg_X x$, to be the minimum of the dimensions of all the simplices that are spanned by some vertices of X and contain x. Since $\mathcal{T}^{U^*}$ is not strongly separable, by Lemma 1, we can find a point $x^* \in \mathcal{CH}(T_R^{U^*}) \cap \mathcal{CH}(T_B^{U^*})$. Let $C_1 = \mathcal{CH}(T_R^{U^*})$ and $C_2 = \mathcal{CH}(T_B^{U^*})$. We claim that $\deg_{C_1} x^* + \deg_{C_2} x^* \geq d - 2$. According to the definition of degree, we can find $\deg_{C_1} x^* + 1$ (resp., $\deg_{C_2} x^* + 1$) points in $T_R^{U^*}$ (resp., $T_B^{U^*}$) such that the simplex spanned by these points, say $\bar{s}_R$ (resp., $\bar{s}_B$), contains x^* in its interior. Let $g : \mathbb{R}^d \to U^*$ and $g' : U^* \to \mathbb{R}^{d-2}$ be the orthogonal projection functions. Clearly, we have $\phi_J = g' \circ g$ for $J = \{3, \ldots, d\}$. Then the convex hull of the g'-images of the vertices of $\bar{s}_R$ (resp., $\bar{s}_B$) contains the point $g'(x^*)$. The g'-images of the points in T^{U^*} are just the points in $\Phi_J(T)$. If $\deg_{C_1} x^* + \deg_{C_2} x^* < d - 2$, we can always find two simplices with the vertices in $\Phi_J(T)$ such that they intersect at $\beta(x^*)$ and the sum of their dimensions is less than $d - 2$. This contradicts the fact that $\Phi_J(T)$ is in general position (note that T has SGPP by our assumption). Thus, $\deg_{C_1} x^* + \deg_{C_2} x^* \geq d - 2$. Now let h be a weak separator of $\mathcal{T}^{U^*}$. Since $x^* \in C_1 \cap C_2$, x^* must be on h. Note that x^* is in the interiors of $\bar{s}_R$ and $\bar{s}_B$. This implies that h must go through all of the $\deg_{C_1} x^* + \deg_{C_2} x^* + 2$ vertices of $\bar{s}_R$ and $\bar{s}_B$. Since $\deg_{C_1} x^* + \deg_{C_2} x^* + 2 \geq d$, and T is in general position, the weak separator h is unique and goes through exactly d points in T^{U^*} (of which at least one is in $T_R^{U^*}$ and one is in $T_B^{U^*}$). □

Let h^* be the unique weak separator of $\mathcal{T}^{U^*}$ described in Lemma 4. We define the extreme separator of $\mathcal{T}$ as the derived separator of h^* in $\mathbb{R}^d$ (see Figure 2.2). At the same time, we call U^* the *auxiliary subspace* defining the extreme separator. Clearly, the extreme separator and the auxiliary subspace are perpendicular to each other.

To once again understand the intuition for the extreme separator, let us consider the case $d = 3$. Imagine there is a plane rotating clockwise around the z-axis.

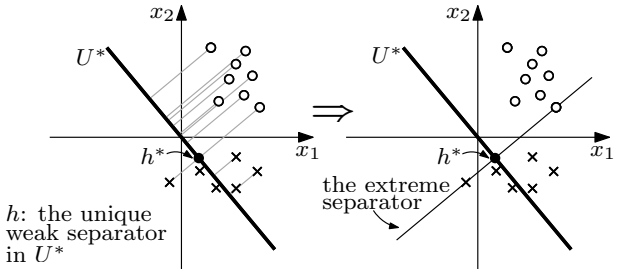


Figure 2.2: Illustrating the extreme separator in $\mathbb{R}^2$.

We keep projecting the points in T (orthogonally) to that plane and track the separability of the projection images. If the images are always separable, then the extreme separator is not defined. Otherwise, there is a closed period of time in which the images are inseparable, which is subsequently followed by an open period in which the images are separable. At the junction of the two periods (from the inseparable one to the separable one), the images are weakly separable by a unique weak separator. Then the rotating plane at this point is just the auxiliary subspace, and the extreme separator is obtained by orthogonally "extending" the unique weak separator to $\mathbb{R}^3$.

2.2.2 Computing the separable-probability

Set $J = \{3, \ldots, d\}$. Consider a realization $\mathcal{T} = (T_R, T_B)$ of $\mathcal{S}$. If $\mathcal{T}$ is separable, then there are two cases: **(1)** the extreme separator of $\mathcal{T}$ is not defined and **(2)** the extreme separator of $\mathcal{T}$ is some hyperplane in $\mathbb{R}^d$. The SP of $\mathcal{S}$ is clearly equal to the sum of the corresponding probabilities of the two cases. By applying Lemma 3, the probability of the first case is equal to the SP of $\Phi_J(\mathcal{S})$, i.e., the bichromatic stochastic dataset obtained by projecting $\mathcal{S}$ via Φ_J (the existence probabilities preserve after projection). On the other hand, if the extreme separator is defined, it must go through exactly d points (of which at least one is in S_R and one is S_B) according to Lemma 4. Thus, the SP of $\mathcal{S}$ can be computed as

$$SP(\mathcal{S}) = SP(\Phi_J(\mathcal{S})) + \sum_{h \in H_S} \tau_{\mathcal{S}}(h),$$

where H_S is the set of the hyperplanes that go through exactly d points in S (of which at least one is in S_R and one is S_B) and $\tau_{\mathcal{S}}(h)$ is the probability that h is the

extreme separator of a realization of $\mathcal{S}$.

To compute $SP(\Phi_J(\mathcal{S}))$ is equivalent to solving the SP problem in $\mathbb{R}^{d-2}$. So it suffices to consider how to compute $\tau_{\mathcal{S}}(h)$ for all $h \in H_S$. Let $h \in H_S$ be a hyperplane and $\mathcal{T} = (T_R, T_B)$ be a realization of $\mathcal{S}$. Clearly, there is a unique element $U^* \in \mathcal{V}$ perpendicular to h (note that all hyperplanes in H_S are not parallel to the x_1x_2-plane due to the SGPP of S). If h is the extreme separator of $\mathcal{T}$, then U^* must be the corresponding auxiliary subspace. Let $E = E_R \cup E_B$ be the set of the d points on h (where $E_R \subseteq S_R$ and $E_B \subseteq S_B$). We investigate the conditions for h to be the extreme separator of $\mathcal{T}$. First, we must have $E \subseteq T_R \cup T_B$. Second, because $\mathcal{T}$ should be weakly (but not strongly) separable after being projected to U^*, there must exist $\hat{r} \in \mathcal{CH}(E_R)$ and $\hat{b} \in \mathcal{CH}(E_B)$ whose projection images on U^* coincide, according to Lemma 1 (actually, such $\hat{r}$ and $\hat{b}$ are unique if they exist, due to the SGPP of S). Finally, since the extreme separator should weakly separate the existent points, all the points in T_R must lie on one side of h while all the points in T_B must lie on the other side, except the points in E. Also, the sides for T_R and T_B are specific, as $\sigma(U^*)$ must be the *clockwise* boundary of P_0. To distinguish the two sides, we define, based on the points $\hat{r}$ and $\hat{b}$, an indicator $o = (o^{(1)}, \ldots, o^{(d)})$, where

$$o^{(1)} = \hat{r}^{(1)} + (\hat{b}^{(2)} - \hat{r}^{(2)}),$$
$$o^{(2)} = \hat{r}^{(2)} + (\hat{r}^{(1)} - \hat{b}^{(1)}),$$
$$o^{(i)} = \hat{r}^{(i)} = \hat{b}^{(i)} \text{ for all } j \in J.$$

(See Figure 2.3 for the location of o.) It is easy to see that, when all the points in T_R (resp., T_B) appear on the same (resp., opposite) side of h with respect to o, $\sigma(U^*)$ is the *clockwise* boundary of P_0. Therefore, we can summarize that h is the extreme separator of a realization R iff

(i) R contains all the points in E;

(ii) there are $\hat{r} \in \mathcal{CH}(E_R)$ and $\hat{b} \in \mathcal{CH}(E_B)$ such that their projection images on U^* coincide;

(iii) R contains no point in S_R (resp., S_B) that is on the opposite (resp., same) side of h with respect to o.

Among the three conditions, the second one has nothing to do with the realization R and can be verified in constant time. If h violates this condition, then

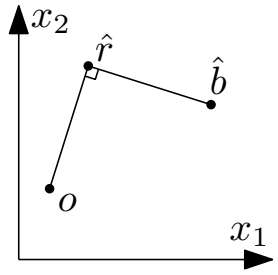


Figure 2.3: Illustrating the location of o. The space in the figure is the 2-dim subspace of $\mathbb{R}^d$ that is parallel to the x_1x_2-plane and contains $\hat{r}$, $\hat{b}$.

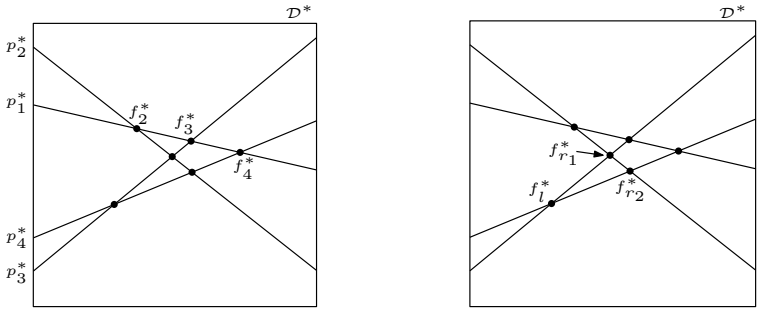
$\tau_{\mathcal{S}}(h) = 0$. Otherwise, $\tau_{\mathcal{S}}(h)$ is just equal to the product of the existence probabilities of the points in E and the non-existence probabilities of the points that R should not contain due to condition **(iii)**. The simplest way to compute it is to scan every point in S once, which takes linear time. This results in an $O(nN^d)$ overall time for computing $SP(\mathcal{S})$, since $|H_S| = O(nN^{d-1})$.

2.2.3 Improving the SP algorithm

To improve the running time of the above algorithm, we can apply the idea of *radial-order* sort in [7]. Specifically, when enumerating the hyperplanes spanned by d points, we first determine $d-1$ points and sort, in $O(N \log N)$ time, all the remaining points around the $(d-2)$-dim subspace spanned by the those $d-1$ points (similar to polar-angle sorting around a point in $\mathbb{R}^2$). Then we consider the last point in that sorted order and maintain a sliding window on the sorted list to record the points on one side of the current hyperplane. In this way, each $\tau_{\mathcal{S}}(h)$ can be computed in amortized constant time by modifying the previous result computed. The time complexity is then reduced to $O(nN^{d-1} \log N)$.

Inspired by [21], we can further improve the algorithm by taking advantage of *duality* [1] and *topological sweep* [32] as follows. We first enumerate $d-2$ points (of which at least one is red and at least one is blue), and these points span a $(d-3)$-dim subspace $\mathcal{D}$, corresponding to a 2-dim dual subspace $\mathcal{D}^*$. By duality, each remaining point p maps to a $(d-1)$-dim hyperplane p^* in the dual space, whose intersection with $\mathcal{D}^*$ is a line l. (Since there is a clear one-to-one correspondence between p^* and l, with a slight abuse of notation, we use p^* to represent l below.) It then follows

that there are $n + N - d + 2 = O(N)$ lines in $\mathcal{D}^*$, forming a line arrangement, and the dual of each intersection point f^* formed by two lines p_1^* and p_2^* is the span f of some $(d-1)$-dim facet in the primal space. We define the *statistic* of f^* as a tuple of the form $(\mathcal{R}^-, \mathcal{R}^+, \mathcal{B}^-, \mathcal{B}^+, \mathcal{T})$, where $\mathcal{R}^-$ and $\mathcal{R}^+$ (resp., $\mathcal{B}^-$ and $\mathcal{B}^+$) denote the product of the non-existence probabilities of the remaining red (resp., blue) points on either side of f, and $\mathcal{T}$ is the set of all the points on f. Given the statistic for f^*, the probability for f^* can be computed in constant time. Thus, it suffices to show how to compute the statistics for all f^* efficiently.



(a) An example of the arrangement in $\mathcal{D}^*$

(b) A elementary step in topological sweep

Figure 2.4: Illustrating how to use duality and topological sweep to eliminate the log factor in runtime.

Assume the lines in $\mathcal{D}$ are $p_1^*, \ldots, p_m^*$, and the intersection points on p_1^* are $f_2^*, \ldots, f_m^*$. Without loss of generality, assume $f_2^*, \ldots, f_m^*$ are sorted from left to right in $\mathcal{D}^*$. We first compute the statistic for f_2^* by brute-force, which takes $O(N)$ time. Then, we move through $f_3^*, \ldots, f_m^*$ in order (see Figure 2.4(a) for an illustration). By duality, the movement from f_{i-1}^* to f_i^* corresponds to the hyperplane rotation from f_{i-1} to f_i with respect to the dual of the line p_1^*, which is a $(d-2)$-dim subspace in the primal space. More importantly, the rotation does not hit any other points except the two points corresponding to p_{i-1}^* and p_i^*. In this way, the statistics of all the intersections along p_1^* can be computed in $O(N)$ time without considering the sorting.

In fact, we cannot afford to sort the intersections on each line since that will take $O(N^2 \log N)$ time. Instead, we compute the entire line arrangement using $O(N^2)$ time and space, then we can visit the intersections on each line in the correct

order (though not necessarily consecutively). To further reduce the space from $O(N^2)$ to $O(N)$, one can perform a *topological sweep* on the arrangement [32]. The topological sweep maintains a cut of size $O(N)$, and sweeps it from left to right over the entire line arrangement using $O(N^2)$ so-called elementary steps, each taking $O(1)$ amortized time (see Figure 2.4(b) for details). Based on this, we find the leftmost intersection point f_l^* in $\mathcal{D}^*$, and compute its statistic by brute-force. This step takes $O(N^2)$ time. Afterwards, when an elementary step is triggered, the statistic for the current intersection point, p^*, can be reported, and we can compute, in $O(1)$ time, the statistics for two more intersections points (e.g., $f_{r_1}^*$ and $f_{r_2}^*$ in Figure 2.4(b)) for future reporting. Thus, as we advance from the leftmost cut to the rightmost one, the statistics of all the intersection points are reported on the fly. Therefore, the runtime of our algorithm is improved to $O(nN^{d-3} \cdot N^2) = O(nN^{d-1})$, using linear space.

Remark. Note that, in $\mathbb{R}^2$ only, the above method actually runs in $O(N^2)$ instead of $O(nN)$. However, the runtime of our previous method based on radial-order sort still remains $O(nN \log N)$.

Theorem 5. *The separable-probability of $\mathcal{S}$ can be computed in $O(nN^{d-1})$ time for $d \geq 3$ and in $O(\min\{nN \log N, N^2\})$ time for $d = 2$.*

2.2.3.1 Application to the SCH membership probability problem

In this section, we give a new reduction from the SCH membership probability problem to the SP problem. By plugging in our SP algorithm presented before, we then obtain a new algorithm for computing the SCH membership probability.

The SCH membership probability problem was introduced for the first time in [7]. The problem can be described as follows. Given a stochastic dataset $\mathcal{S}$ in $\mathbb{R}^d$ and a query point $q \in \mathbb{R}^d$, compute the probability that q is inside a SCH of $\mathcal{S}$, which we call the *SCH membership probability* (SCHMP) of q with respect to $\mathcal{S}$.

It has been shown in [21] that one can reduce the SCHMP problem in $\mathbb{R}^d$ to the SP problem in $\mathbb{R}^{d-1}$. Here, we provide a more direct and simpler reduction. Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset and q be a query point. Set $m = |S|$. Clearly, q is outside the SCH of $\mathcal{S}$ iff it can be separated from the realization of $\mathcal{S}$ by a hyperplane. Thus, we construct a bichromatic stochastic dataset $\mathcal{S}' = (S_R, S_B, \pi')$,

where $S_R = \{q\}$, $S_B = S$, and

$$\pi'(a) = \begin{cases} 1 & \text{if } a = q, \\ \pi(a) & \text{otherwise.} \end{cases}$$

Then the SCHMP of q with respect to $\mathcal{S}$ is just equal to $1 - SP(\mathcal{S}')$. It can be computed in $O(m^{d-1})$ time for $d \geq 3$ and $O(m \log m)$ time for $d = 2$ by applying our SP algorithm (since $|S_R| = 1$ and $|S_B| = m$), matching the time bound in [21].

Theorem 6. *One can compute the SCH membership probability of a point with respect to a stochastic dataset of size m in $\mathbb{R}^d$ in $O(m^{d-1})$ time.*

Interestingly, this method for computing SCHMP is a generalization of the witness-edge method in [7] to the case $d > 2$, where the latter was the first known approach that solves this problem in $\mathbb{R}^2$ and was thought to be difficult to be generalized to higher dimensions [7]. This can be seen as follows. When plugging in our SP algorithm, we enumerate all the possible extreme separators of $\{q\} \cup R$, where R is a realization of $\mathcal{S}$. The extreme separator goes through d points, in which one is q. These d points corresponds to a facet of the convex polytope $\mathcal{CH}(\{q\} \cup R)$ adjacent to q. This facet is uniquely determined by $\mathcal{CH}(\{q\} \cup R)$. We call it the *witness-facet* of q in $\mathcal{CH}(\{q\} \cup R)$. Then enumerating the possible extreme separators is equivalent to enumerating the possible witness-facet of q in $\mathcal{CH}(\{q\} \cup R)$. When $d = 2$, the notion of witness-facet coincides with the notion of *witness-edge* defined in [7]. Thus, our method is identical to the witness-edge method in $\mathbb{R}^2$, and both methods have the same $O(m \log m)$ runtime. For $d \geq 3$, a different method was given in [7] for computing SCHMP, whose time complexity is $O(m^d)$. In this case, our $O(m^{d-1})$-time algorithm improves the bound by a factor of m.

2.2.4 Witness-based lower bound for separable-probability

When solving the SP problem, the key idea of our algorithm is to group the probabilities of those separable realizations which share the same extreme separator so that the SP can be efficiently computed by considering the extreme separators instead of single realizations. By extending and abstracting this idea, we are able to get a general framework for computing SP, which we call the *witness-based framework*. Let $\mathcal{S}$ be the given stochastic dataset and $\mathcal{I}_{\mathcal{S}}$ be the set of all the separable

realizations of $\mathcal{S}$. The witness-based framework for computing the SP of $\mathcal{S}$ is the following. Here $\mathcal{P}(\cdot)$ denotes the power set.

1. Define a set $W = \{h_1, \ldots, h_m\}$ of hyperplanes (called *witness separators*) with specified weights $w_1, \ldots, w_m$ and an implicitly specified witness rule $f : W \rightarrow \mathcal{P}(\mathcal{I}_S)$ such that

   - the elements in $f(h_i)$ are (either strongly or weakly) separated by h_i;
   - the witness probability (see Step **2** below) of each h_i is efficiently computable;
   - any element $I \in \mathcal{I}_S$ satisfies $\sum_{\forall i(I \in f(h_i))} w_i = 1$.

   We say the witness separator h_i *witnesses* the elements in $f(h_i)$.

2. Compute **efficiently** the *witness probability* of each $h_i \in W$, which is defined as
$$witP(h_i) = \sum_{I \in f(h_i)} Pr(I),$$
where $Pr(I)$ is the probability that I is a realization of $\mathcal{S}$.

3. Compute $SP(\mathcal{S})$ by linearly combining the witness probabilities with the specified weights, i.e.,
$$Sep(S) = \sum_{i=1}^{m} (w_i \cdot witP(h_i)) = \sum_{I \in \mathcal{I}_S} Pr(I).$$

Note that the witness-based framework is very general. The ways of defining witness separators and specifying witness rules may vary among different witness-based algorithms. Our algorithm and the one introduced in [21], which are the only two known algorithms for computing SP at this time, both belong to the witness-based framework. Similar frameworks are also used to solve other probability-computing problems. For example, the two algorithms in [7] for computing convex hull membership probability are both implemented by defining witness edges/facets and summing up the witness probabilities. To the best of our knowledge, up to now, most probability-computing problems for geometric uncertain datasets are solved by applying ideas close to this framework.

Now we show that any SP computing algorithm following the witness-based framework takes at least $\Omega(nN^{d-1})$ time in the worst case, and thus our algorithm is optimal among this category of algorithms for any $d \geq 3$. Clearly, the runtime of a witness-based algorithm is at least $|W| = m$, i.e., the number of the witness separators. Then a question naturally arises: how many witness separators do we need for computing SP? From the above framework, one restriction for W is that each separable instance of S must be witnessed by at least one witness separator $h_i \in W$, i.e., $\mathcal{I}_S = \bigcup_{i=1}^{m} f(h_i)$. Otherwise, the probabilities of the unwitnessed instances in $\mathcal{I}_S$ will not be counted when computing $SP(\mathcal{S})$. It then follows that each separable realization of $\mathcal{S}$ must be separated by some $h_i \in W$. We prove that, in the worst case, we always need $\Omega(nN^{d-1})$ hyperplanes to separate all the separable realizations of $\mathcal{S}$, which implies an $\Omega(nN^{d-1})$ lower bound on the runtime of any witness-based SP computing algorithm. We say a hyperplane set H *covers* a bichromatic dataset $\mathcal{T} = (T_R, T_B)$ iff for any non-trivial separable subset $\mathcal{V} \subseteq \mathcal{T}$ (i.e., $\mathcal{V}$ contains at least one red point and one blue point), there exists $h \in H$ that separates $\mathcal{V}$. We define $\chi(\mathcal{T})$ as the cardinality of the smallest set of hyperplanes that cover $\mathcal{T}$. The following theorem completes the discussion, and is also of independent interest.

Theorem 7. *Let $\mathcal{D}^d_{n,N}$ be the collection of all the bichromatic datasets in $\mathbb{R}^d$ of size (n, N). Define*

$$\Gamma_d(n, N) = \sup_{\mathcal{T} \in \mathcal{D}^d_{n,N}} \chi(\mathcal{T}).$$

Then for all constant d, we have $\Gamma_d(n, N) = \Omega(nN^{d-1})$.

Proof. To prove this theorem, it is more convenient to work on "directed" hyperplanes. A *directed hyperplane* in $\mathbb{R}^d$ is a hyperplane with one side (half-space) specified to be red and the other side specified to be blue. It can be represented as a $(d+1)$-tuple $(a_0, a_1, \ldots, a_d)$ of real numbers (not all equal to 0 simultaneously) such that the inequality $a_0 + \sum_{i=1}^{d} a_i x_i < 0$ indicates the red side. We say the directed hyperplane $(a_0, a_1, \ldots, a_d)$ *separates* a bichromatic dateset $\mathcal{T} = (T_R, T_B)$ iff there is no point located on the side of different color, i.e., for each point $x = (x_1, \ldots, x_d) \in T$

where $T = T_R \cup T_B$, we have

$$a_0 + \sum_{i=1}^{d} a_i x_i \begin{cases} \leq 0 & \text{if } x \in T_R, \\ \geq 0 & \text{if } x \in T_B. \end{cases}$$

Since a (undirected) hyperplane can be replaced with two directed hyperplanes, the number of the directed hyperplanes required for covering a dataset is at most twice the number of the undirected ones. Thus, it suffices to prove the result with respect to directed hyperplanes. In the rest of the proof, the notation $\chi(\mathcal{T})$ is used to denote the size of the smallest set of directed hyperplanes (instead of hyperplanes) which cover $\mathcal{T}$.

We show that, for all constant d, there exists some bichromatic dataset $\mathcal{T} \in \mathcal{D}^d_{n,N}$ with general position such that $\chi(\mathcal{T}) = \Omega(nN^{d-1})$. Specifically, we use induction on the dimension d. The base case $d = 1$ is trivial. Assume the argument holds for $d = k - 1$, and we consider the case of $d = k$. We want to construct a bichromatic dataset $\mathcal{T}$ in $\mathbb{R}^k$ of size (n, N) such that $\chi(\mathcal{T}) = \Omega(nN^{k-1})$.

Our first step is to construct a bichromatic dataset $\mathcal{T}'$ in $\mathbb{R}^k$ of size $(1, N)$ such that $\chi(\mathcal{T})' = \Omega(N^{k-1})$. By our induction hypothesis, there exists a bichromatic dataset $\mathcal{U} = (U_R, U_B)$ in $\mathbb{R}^{k-1}$ (in general position) of size (N, N) such that $\chi(\mathcal{U}) = \Omega(N^{k-1})$. Define two functions $f_R, f_B : \mathbb{R}^{k-1} \to \mathbb{R}^k$ as

$$f_R : (x_1, \ldots, x_{k-1}) \mapsto (-x_1, \ldots, -x_{k-1}, -1),$$

$$f_B : (x_1, \ldots, x_{k-1}) \mapsto (x_1, \ldots, x_{k-1}, 1).$$

Let r be the origin of $\mathbb{R}^k$. We then define $\mathcal{T}' = (T'_R, T'_B)$ where $T'_R = \{r\}$ and $T'_B = f_R(U_R) \cup f_B(U_B)$. We claim that $\chi(\mathcal{T}') \geq \chi(\mathcal{U})$, which implies $\chi(\mathcal{T}') = \Omega(N^{k-1})$. For any nontrivial separable subset $\mathcal{V} = (V_R, V_B) \subseteq \mathcal{U}$, define $f(\mathcal{V}) = (\{r\}, f_R(V_R) \cup f_B(V_B))$ as a bichromatic dataset in $\mathbb{R}^k$. It is easy to see that $\mathcal{V}$ is separable iff $f(\mathcal{V})$ is. Indeed, if a non-horizontal (i.e., not parallel to the plane $x_k = 0$) directed hyperplane $(a_0, a_1, \ldots, a_k)$ in $\mathbb{R}^k$ separates $f(\mathcal{V})$, then we have a corresponding directed hyperplane $(a_0 + a_k, a_1, \ldots, a_{k-1})$ in $\mathbb{R}^{k-1}$ that separates $\mathcal{V}$. We call the latter the *induced* plane of the former. Now let $H = \{h_1, \ldots, h_{\chi(T')}\}$ be a set of directed hyperplanes in $\mathbb{R}^k$ which cover $\mathcal{T}'$. Assume they are all non-horizontal (if any of them is horizontal, we can always slightly rotate it without changing the subsets of $\mathcal{T}'$ it separates). Then let $H' = \{h'_1, \ldots, h'_{\chi(T')}\}$ be a set of

directed hyperplanes in $\mathbb{R}^{k-1}$ in which h'_i is the induced plane of h_i. Then H' covers $\mathcal{U}$, which implies that $\chi(\mathcal{U}) \leq \chi(\mathcal{T}')$.

The next step is to extend $\mathcal{T}'$ into another set $\mathcal{T}$ of size (n, N) in $\mathbb{R}^k$ such that $\chi(\mathcal{T}) = \Omega(nN^{k-1})$. Recall that r is the only point in T'_R, which is the origin of $\mathbb{R}^k$. We denote by $b_1, \ldots, b_{2N}$ the $2N$ points in T'_B. We first slightly perturb each b_i without changing $\chi(\mathcal{T}')$ to make the points $r, b_1, \ldots, b_{2N}$ in general position. For convenience, we now use $\mathcal{T}'$ to denote the new dataset after the perturbation. Then we find an ε-ball centered at the origin of $\mathbb{R}^k$ with a sufficiently small $\varepsilon > 0$ such that if the point r perturbs inside that ball, $\chi(\mathcal{T}')$ does not change. The value of ε can be determined as follows. For each $(k-1)$-dim linear subspace spanned by k points $b_{\pi_1}, \ldots, b_{\pi_k}$, we compute the distance from the origin to it. Then we set ε to be a number less than the minimum of those distances. Inside this ε-ball, we pick n points $r_1, \ldots, r_n$ such that all the points $r_1, \ldots, r_n, b_1, \ldots, b_{2N}$ are in general position. Define $T_R = \{r_1, \ldots, r_n\}$. Next, we find another small number $\varepsilon' > 0$ such that for any hyperplane h in $\mathbb{R}^k$, there are at most k points among $r_1, \ldots, r_n$ whose distances to h are less than or equal to ε'. We can determine ε' as follows. For each $(k+1)$-tuple $t = (r_{\pi_1}, \ldots, r_{\pi_{k+1}})$, we define

$$\delta_t = \inf_h \max_{i=1}^{k+1} dist(h, r_{\pi_i}).$$

The we set ε' to be a number less than the minimum of all δ_t. Clearly, ε' satisfies the desired property. Now, for each r_i, we find $k+1$ points $b'_{i,1}, \ldots, b'_{i,k+1}$ inside the ε'-ball centered at r_i such that the simplex spanned by $b'_{i,1}, \ldots, b'_{i,k+1}$ contains r_i in its interior. We carefully determine the locations of these points to guarantee the general-position property. Then we define T_B as the set consisting of $b_1, \ldots, b_{2N}$ and all $b'_{i,j}$ for $i \in \{1, \ldots, n\}$ and $j = \{1, \ldots, k+1\}$. Set $\mathcal{T} = (T_R, T_B)$, which is of size $(n, 2N + (k+1)n)$. We show that $\chi(\mathcal{T}) = \Omega(nN^{k-1})$. Let H be any set of directed hyperplanes which cover $\mathcal{T}$. Also, let $H_i \subseteq H$ be the subset of the directed hyperplanes whose distances to the point r_i are at most ε'. We claim that $|H_i| \geq \chi(\mathcal{T}')$ for all $i \in \{1, \ldots, n\}$. Set $\mathcal{T}'' = (T''_R, T''_B)$ where $T''_R = \{r_i\}$ and $T''_B = \{b_1, \ldots, b_{2N}\}$. Recall that r_i is inside the ε-ball centered at the origin of $\mathbb{R}^k$, which implies $\chi(\mathcal{T}'') = \chi(\mathcal{T}')$. Assume that $|H_i| < \chi(\mathcal{T}'')$. Then H_i does not cover $\mathcal{T}''$. Let $\mathcal{V} \subseteq \mathcal{T}''$ be a nontrivial separable subset that is not separated by any $h \in H_i$. Let h^* be a directed hyperplane which goes through r_i and weakly separates $\mathcal{V}$.

Consider the points $b'_{i,1}, \ldots, b'_{i,k+1}$. Since r_i is in the interior of the simplex spanned by $b'_{i,1}, \ldots, b'_{i,k+1}$, we can find at least one point $b'_{i,j}$ such that $(V_R, V_B \cup \{b'_{i,j}\})$ is also separated by h^* (and thus separable). We show that $(V_R, V_B \cup \{b'_{i,j}\})$ is not separated by any $h \in H$, which contradicts the fact that H covers $\mathcal{T}$. We consider two cases: $h \in H_i$ and $h \in H \backslash H_i$. Any $h \in H_i$ is not a separator of $(V_R, V_B \cup \{b'_{i,j}\})$ because it does not separate $\mathcal{V}$. For any $h \in H \backslash H_i$, we notice that $dist(h, r_i) > \epsilon'$. Thus, both r_i and $b'_{i,j}$ are on the same side of h, which implies that h is not a separator of $(V_R, V_B \cup \{b'_{i,j}\})$. As a result, we have $|H_i| \geq \chi(\mathcal{T}'') = \chi(\mathcal{T}')$. Now recall that for any hyperplane h in $\mathbb{R}^k$, there are at most k points among $r_1, \ldots, r_n$ whose distances to h are less than or equal to ε'. This implies that

$$|H| \geq \sum_{i=1}^{n} \frac{|H_i|}{k} \geq \frac{n\chi(\mathcal{T}')}{k}.$$

Therefore, we know that $\chi(\mathcal{T})$ is $\Omega(nN^{k-1})$. Note that the size of $\mathcal{T}$ is now $(n, 2N + (k+1)n)$. To make it exactly (n, N), we only need to choose $n_0 = n/(3k+3)$ and $N_0 = N/3$, and use the same method to construct a bichromatic dataset $\mathcal{T}$ of size $(n_0, 2N_0 + (k+1)n_0)$ in general position such that $\chi(\mathcal{T}) = \Omega(n_0 N_0^{k-1}) = \Omega(nN^{k-1})$. Then by adding some dummy points, we eventually obtain $\mathcal{T} \in \mathcal{D}^d_{n,N}$ with $\chi(\mathcal{T}) = \Omega(nN^{k-1})$. □

2.3 Expected separation-margin

We study the problem of computing the expected separation-margin (ESM) of $\mathcal{S}$ (denoted by $ESM(\mathcal{S})$), i.e., the expectation of the separation-margin of a realization of $\mathcal{S}$ (which was defined formally in Section 2.1). We begin by introducing some notions. Let $\mathcal{T} = (T_R, T_B)$ be a nontrivial bichromatic dataset.

Lemma 8. *There exists a unique maximum-margin separator h of $\mathcal{T}$. Furthermore, for any closest pair (r, b) of points where $r \in \mathcal{CH}(T_R)$ and $b \in \mathcal{CH}(T_B)$, h is he bisector of the segment $[r, b]$ connecting r and b.*

Proof. Let (r, b) be any closest pair of points where $r \in \mathcal{CH}(T_R)$ and $b \in \mathcal{CH}(T_B)$. Also, let h be the bisector of the segment $[r, b]$. Then $M_h(\mathcal{T}) = \text{dist}(r, b)/2$. Let $h' \neq h$ be another separator of $\mathcal{T}$. We have that

$$\min\{\text{dist}(r, h'), \text{dist}(b, h')\} < \text{dist}(r, b)/2.$$

Furthermore, since $r \in \mathcal{CH}(T_R)$ and $b \in \mathcal{CH}(T_B)$, $M_{h'}(\mathcal{T})$ must be less than or equal to $\min\{\text{dist}(r, h'), \text{dist}(b, h')\}$. Therefore,

$$M_{h'}(\mathcal{T}) \leq \min\{\text{dist}(r, h'), \text{dist}(b, h')\} < \text{dist}(r, b)/2 = M_h(\mathcal{T}).$$

So h' is not a maximum-margin separator of $\mathcal{T}$. It follows that h is the unique maximum-margin separator of $\mathcal{T}$, though the closest pair (r, b) may be not unique. □

Let h be the maximum-margin separator of $\mathcal{T}$ and $M = Mar(\mathcal{T})$ be its separation-margin. Define $C_R = \{r \in T_R : \text{dist}(r, h) = M\}$ and $C_B = \{b \in T_B : \text{dist}(b, h) = M\}$. We define $\text{Supp}(\mathcal{T}) = (C_R, C_B)$ and call this the *support set* of $\mathcal{T}$. Note that all the points in $C_R \cup C_B$ have the same distance to h. Thus, there exist two parallel hyperplanes h_r and h_b (both of which are parallel to h) where h_r goes through all the points in C_R and h_b goes through all the points in C_B. We call h_r and h_b the *support planes* of $\mathcal{T}$. Including the maximum-margin separator h, they form a group of three parallel and equidistant hyperplanes (h_r, h, h_b) (see Figure 2.5). Since the maximum-margin separator is unique, the support set and support planes are also unique. We shall show that the maximum-margin separator can be uniquely determined via the support set.

Lemma 9. *Let $\mathcal{C} = (C_R, C_B) = Supp(\mathcal{T})$. Then $\mathcal{T}$ and $\mathcal{C}$ share the same maximum-margin separator as well as the same separation-margin. Furthermore, $Supp(\mathcal{C}) = \mathcal{C}$.*

Proof. Let h be the maximum-margin separator of $\mathcal{T}$ and M be the separation-margin of $\mathcal{T}$. Also, let (r, b) be any closest pair of points where $r \in \mathcal{CH}(T_R)$ and $b \in \mathcal{CH}(T_B)$. From the proof of Lemma 8, we know that $\text{dist}(r, h) = \text{dist}(b, h) = M$. It immediately follows that $r \in \mathcal{CH}(C_R)$ and $b \in \mathcal{CH}(C_B)$. Since $\mathcal{CH}(C_R) \subseteq \mathcal{CH}(T_R)$ and $\mathcal{CH}(C_B) \subseteq \mathcal{CH}(T_B)$, (r, b) is also a closest pair of points for $r \in \mathcal{CH}(C_R)$ and $b \in \mathcal{CH}(C_B)$. Thus, h is also the maximum-margin separator of $\mathcal{C}$, and the separation-margin of $\mathcal{C}$ is equal to that of $\mathcal{T}$. Furthermore, because all of the points in $\mathcal{C}$ have the same distance to h, we have $\text{Supp}(\mathcal{C}) = \mathcal{C}$. □

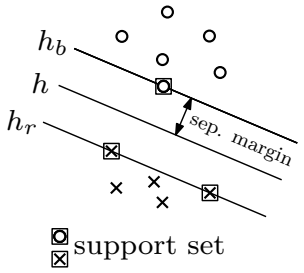


Figure 2.5: An example of support set and support plane, in $\mathbb{R}^2$

2.3.1 Computing the expected separation-margin

According to Lemma 9, the separation-margin of a separable bichromatic dataset is equal to that of its support set. Thus, the ESM of $\mathcal{S}$ can be computed as

$$ESM(\mathcal{S}) = \sum_{\mathcal{C}} \xi_{\mathcal{S}}(\mathcal{C}) \cdot Mar(\mathcal{C}), \tag{2.1}$$

where $\xi_{\mathcal{S}}(\mathcal{C})$ is the probability that a realization of $\mathcal{S}$ is separable with the support set $\mathcal{C}$. Since S has the general position property, the size of the support set of a separable realization of $\mathcal{S}$ can be at most $2d$ (d points in S_R and d points in S_B at most). It follows that the total number of the possible support sets to be considered is bounded by $O(n^d N^d)$. Indeed, we can further improve this bound.

Lemma 10. *The total number of the possible support sets of the realizations of $\mathcal{S}$ is $O(nN^d)$. As a result, the number of the (distinct) possible separation-margins is also bounded by $O(nN^d)$.*

Proof. The number of possible support sets of size smaller than or equal to d is clearly bounded by $O(nN^d)$. So we only need to bound the number of the ones of sizes are larger than d. We first arbitrarily label all the points in S from 1 to $n+N$. For any subset $(C_R, C_B) \subseteq \mathcal{S}$ with $|C_R \cup C_B| > d$, define the representation of (C_R, C_B) as the set of the $d+1$ points in $C_R \cup C_B$ with the smallest labels. Let $\{a_1, \ldots, a_{d+1}\} \subseteq S$ be a subset of $d+1$ points, where $a_1, \ldots, a_k \in S_R$ and $a_{k+1}, \ldots, a_{d+1} \in S_B$. We consider the possible support sets whose representation is $\{a_1, \ldots, a_{d+1}\}$. If $k = 0$ or $k = d+1$, there is no possible support set represented by $\{a_1, \ldots, a_{d+1}\}$, because number of the red/blue points in a support set can at most be d. Now consider the case that $1 \leq k \leq d$. It is easy to see that there exists

a unique pair of parallel hyperplanes (h_r, h_b) such that h_r goes through $a_1, \ldots, a_k$ and h_b goes through $a_{k+1}, \ldots, a_{d+1}$, since S is in general position. If (C_R, C_B) is the support set of a separable realization of $\mathcal{S}$ represented by $\{a_1, \ldots, a_{d+1}\}$, then h_r and h_b must be the corresponding support planes of that realization. That means all the points in C_R and C_B must lie on h_r and h_b, respectively. Note that there are at most $2d$ points on $h_r \cup h_b$, which implies that the number of the possible support sets represented by $\{a_1, \ldots, a_{d+1}\}$ is constant. Since the number of such subsets is $O(nN^d)$, $\mathcal{S}$ can have at most $O(nN^d)$ possible support sets. Finally, because the separation-margin is uniquely determined by the support set, the number of the possible separation-margins is also bounded by $O(nN^d)$. □

By applying Equation 2.1, we can enumerate all the $O(nN^d)$ possible support sets to compute the ESM of $ESM(\mathcal{S})$. The $O(nN^d)$ possible support sets can be enumerated as follows. For the ones of sizes less than $d+1$, we enumerate them in the obvious way. For the ones of sizes larger than or equal to $d+1$, we first enumerate a subset $\{a_1, \ldots, a_{d+1}\} \subseteq S$ (in which at least one point is in S_R and one point is in S_B), which would be the representation of the support sets (see the proof of Lemma 10). Via this subset, we can uniquely determine two parallel hyperplanes h_r and h_b where h_r goes through the points in $\{a_1, \ldots, a_{d+1}\} \cap S_R$ and h_b goes through the points in $\{a_1, \ldots, a_{d+1}\} \cap S_B$. We then find all the points on h_r and h_b, the number of which is at most $2d$, including $\{a_1, \ldots, a_{d+1}\}$. Once we have those points, we are able to enumerate all the possible support sets represented by $\{a_1, \ldots, a_{d+1}\}$. For each such possible support set $\mathcal{C} = (C_R, C_B)$, $Mar(\mathcal{C})$ can be straightforwardly computed in constant time since the size of $\mathcal{C}$ is constant. To compute $\xi_{\mathcal{S}}(\mathcal{C})$, we observe that $\mathcal{C}$ is the support set of a realization R of $\mathcal{S}$ iff

1) all the points in C_R (resp., C_B) lie on h_r (resp., h_b);

2) R contains all the points in $C = C_R \cup C_B$;

3) none of the points in S_R (resp., S_B) on the same side of h_r (resp., h_b) as h is contained in R;

4) except the points in C, none of the points in S_R (resp., S_B) on h_r (resp., h_b) is contained in R.

The first condition can be easily verified. If $\mathcal{C}$ violates this condition, then $\xi_{\mathcal{S}}(\mathcal{C}) = 0$. Otherwise, $\xi_{\mathcal{S}}(\mathcal{C})$ is just equal to the product of the existence probabilities of the

points in $\mathcal{C}$ (the second condition) and the non-existence probabilities of those points that should not be contained a realization (the last two conditions). If we use the simplest way, i.e., scanning all the points in S, to find the points on h_r and h_b (for enumerating the possible support sets represented by a set of $d+1$ points) as well as to compute each $\xi_{\mathcal{S}}(\mathcal{C})$, then the total time for computing $ESM(\mathcal{S})$ is $O(nN^{d+1})$.

2.3.2 Improving the ESM algorithm

It is easy to improve the running time of the above algorithm to $O(nN^d \log N)$ by slightly modifying the sort method we used for improving our SP algorithm. When enumerating $d+1$ points in S, we first determine d points (of which at least one is in S_R and one is in S_B). Let $r_1, \ldots, r_k \in S_R$ and $b_1, \ldots, b_{d-k} \in S_B$ denote these d points. We can uniquely determine two parallel $(d-2)$-dim linear subspaces X_r and X_b of $\mathbb{R}^d$ such that $r_1, \ldots, r_k \in X_r$ and $b_1, \ldots, b_{d-k} \in X_b$. We sort all the remaining points in S_R around X_r and those in S_B around X_b. Then we consider the last point in that sorted order (say the ones in S_R first and then those in S_B) and meanwhile maintain two sliding windows (for the points in S_R and S_B respectively). In this way, we are able to use amortized constant time to consider each set of $d+1$ points, i.e., to compute the probabilities of all the possible support sets represented by the $d+1$ points and add the portions contributed by these possible support sets to the ESM. Thus, the computation of ESM can be done in $O(nN^d \log N)$ time.

To further improve the time complexity to $O(nN^d)$ requires more work. We can still apply the duality and topological sweep techniques but the approach is somewhat different from that in the SP problem. For convenience, we simply call the points in S_R (resp., S_B) *red* (resp., *blue*) points. We define the *red* (resp., *blue*) *statistics* of a hyperplane h as a tuple consisting of the set of the red (resp., blue) points on h and the product of the non-existence probabilities of all the red (resp., blue) points on each side of h. As we see, in the process of computing the separable-probability, the object enumerated is one hyperplane spanned by d points and what we want to compute is the red and blue statistics of the hyperplane. In this situation, the idea of duality and topological sweep can be directly used to improve the efficiency of each computation. However, when computing the expected separation-margin, the situation is different. At each step, we have three parallel

and equidistant hyperplanes (h_r, h, h_b) determined by $d+1$ points, and what we want to compute is the red statistics of h_r and the blue statistics of h_b. Thus, in order to apply the duality and topological sweep techniques, our idea is to transform the problem from the latter form to the former one. We consider two different cases: $d \geq 3$ and $d = 2$.

Suppose $d \geq 3$. In this case, when enumerating $d+1$ points, we first determine two of them, of which one is in S_R (say r) and the other is in S_B (say b). Let c be the midpoint of the segment $[r, b]$. Then for each $r_i \in S_R$, we construct a new point $r'_i = r_i + \overrightarrow{rc}$, and for each $b_i \in S_B$, we construct a new point $b'_i = b_i + \overrightarrow{bc}$. We denote by S'_R the set of all r'_i and by S'_B the set of all b'_i. We construct a new bichromatic stochastic dataset $\mathcal{S}' = (S'_R, S'_B, \pi')$ where $\pi'(r'_i) = \pi(r_i)$ and $\pi'(b'_i) = \pi(b_i)$, and set $S' = S'_R \cup S'_B$. Now consider a set of $d+1$ points in S including r and b. Let (h_r, h, h_b) be the three hyperplanes determined by these $d+1$ points. In order to complete the computation, what we need to know is the red statistics of h_r and the blue statistics of h_b. According to the construction of $\mathcal{S}'$, one can easily verify the following facts.

- A point $r_i \in S_R$ (resp., $b_i \in S_B$) is on h_r (resp., h_b) iff its corresponding point $r'_i \in S'_R$ (resp., $b'_i \in S'_B$) is on h. So each of the $d+1$ points corresponds to a point in S' that is on h.

- The points in S_R (resp., S_B) on each side of h_r (resp., h_b) correspond to the points in S'_R (resp., S'_B) on each side of h.

Based on the above observations, the red statistics of h_r and the blue statistics of h_b with respect to $\mathcal{S}$ just correspond to the red and blue statistics of h with respect to $\mathcal{S}'$. In other words, to consider all the possible support sets represented by these $d+1$ points, it suffices to know the red and blue statistics of h with respect to $\mathcal{S}'$. Now the problem we face is similar to that in the SP problem. We want to compute, for each hyperplane h spanned by the point c and other $d-1$ points in S', the red and blue statistics of h. By applying the idea of duality and topological sweep, this can be done in $O(N^{d-1})$ time. This is the runtime for a fixed pair (r, b). To compute the ESM, we need to enumerate all $O(nN)$ such pairs, so the overall time is $O(nN^d)$.

For the case of $d = 2$, however, the above method does not work. Since we enumerate three points when $d = 2$, if we first determine two of them (say a and b), we are not able to create the line arrangement in the dual space and use topological sweep to complete the computation work for the pair (r, b) in $O(N)$ time. So we need to deal with the case of $d = 2$ separately. Without loss of generality, we only consider the case where one of the three points enumerated is in S_R and the other two are in S_B (as the two-red one-blue case is symmetric). Let $n_r = |S_R|$ and $n_b = |S_B|$. When enumerating three points, we first determine a point $r \in S_R$ and sort all the other points in S_R around r; let L be the resulting sorted list. Then for all the points in S_B, we construct their dual lines to form a line arrangement. Each vertex (i.e., intersection point) of the arrangement corresponds to a pair (b_i, b_j) of points in S_B. We want to apply topological sweep on the arrangement and consider each set $\{r, b_i, b_j\}$ of three points at the time we visit the vertex corresponding to (b_i, b_j). Fix $\{r, b_i, b_j\}$, and let (h_r, h, h_b) be the three hyperplanes determined by $\{r, b_i, b_j\}$. In order to complete the computation, we need the red statistics of h_r and the blue statistics of h_b. We note that the hyperplane h_b is actually determined only by b_i and b_j (and independent of r). Thus, the blue statistics of h_b can be directly computed in the process of topological sweep. The crucial part is to compute the red statistics of h_r. What we do is to maintain n_b sliding windows $w_1, \ldots, w_{n_b}$ on the sorted list L, where w_i corresponds to the point b_i. During the topological sweep, the sliding window w_i dynamically indicates the red points on one side of the hyperplane h_r determined by the set $\{r, b_i, b^*\}$, where (b_i, b^*) is the most recently visited vertex on the dual line of b_i. At each time a new vertex (b_i, b_j) is visited, we update w_i and w_j, and meanwhile compute the red statistics of the hyperplane h_r determined by the set $\{r, b_i, b_j\}$. It is easy to see that both updating the sliding windows and computing the statistics can be done in amortized constant time. Therefore, for each red point r, the computations take $O(n_b^2)$ time. The total time for considering all the red points is then $O(n_r n_b^2)$, which is bounded by $O(nN^2)$. Symmetrically, the work for enumerating two red points and one blue point can also be done in $O(nN^2)$ time.

Theorem 11. *The expected separation-margin of $\mathcal{S}$ can be computed in $O(nN^d)$ time.*

2.3.3 Hardness of computing expected separation-margin

We show that the bound achieved in Lemma 10 is tight, which suggests that our algorithm for computing ESM may be difficult to be further improved. For a bichromatic stochastic dataset $\mathcal{S}$, define $\kappa(\mathcal{S})$ as the total number of the possible separation-margins of the realizations of $\mathcal{S}$.

Theorem 12. *For any constant d, there exists some bichromatic stochastic dataset $\mathcal{S}$ in $\mathbb{R}^d$ of size (n, N) such that $\kappa(\mathcal{S}) = \Theta(nN^d)$.*

Proof. First, we construct $(d+1)$ points $c_0, c_1, \ldots, c_d \in \mathbb{R}^d$ as

$$c_0 = (0, \ldots, 0),$$

$$c_i = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{d-i}), \text{ for } i = 1, \ldots, d.$$

Define $B_0, B_1, \ldots, B_d$ as the ε-balls (ε is a sufficiently small positive constant) centered at $c_0, c_1, \ldots, c_d$ respectively. We randomly generate a bichromatic stochastic dataset $\mathcal{S}^* = (S_R^*, S_B^*, \pi)$ with $|S_R^*| = n$ and $|S_B^*| = N$ (where $n \leq N$) as follows. The function π is set to be a constant function assigning all points an identical existence probability equal to 0.5. The points in S_R^* are drawn from the uniform distribution on B_0. For the points in S_B^*, we evenly separate them into d groups each of which contains N/d points (for convenience, assume N is a multiple of d). The points in the i-th group are drawn from the uniform distribution on B_i. All the points are drawn independently.

We show that

$$\Pr\left[\ \kappa(\mathcal{S}^*) \geq n\left(\frac{N}{d}\right)^d\ \right] > 0,$$

which implies the existence of a bichromatic stochastic dataset $\mathcal{S}$ of size (n, N) satisfying $\kappa(\mathcal{S}) = \Theta(nN^d)$. We denote by $r_1, \ldots, r_n$ the n random points in S_R^* and by $b_{i,1}, \ldots, b_{i,N/d}$ the N/d random points in S_B^* that are drawn from B_i for $i \in \{1, \ldots, d\}$. Consider all the $(d+1)$-tuples $(j, \pi_1, \ldots, \pi_d)$ where $j \in \{1, \ldots, n\}$ and $\pi_1, \ldots, \pi_d \in \{1, \ldots, N/d\}$. Clearly, we have in total $n(N/d)^d = \Theta(nN^d)$ such tuples. For each such tuple $\tau = (j, \pi_1, \ldots, \pi_d)$, define M_τ as the separation-margin of $\{r_j, b_{1,\pi_1}, \ldots, b_{d,\pi_d}\}$, which is a real-valued random variable.

We claim that $\Pr[M_\tau = M_{\tau'}] = 0$ for any two distinct tuples τ and τ' (though $M_\tau = M_{\tau'}$ is not an impossible event). Let $\tau = (j, \pi_1, \ldots, \pi_d)$ be a tuple where $j \in \{1, \ldots, n\}$ and $\pi_1, \ldots, \pi_d \in \{1, \ldots, N/d\}$. Let h denote the (probabilistic) hyperplane going through $b_{1,\pi_1}, \ldots, b_{d,\pi_d}$. We first observe that $M_\tau = \mathrm{dist}(r_j, h)/2$. Indeed, due to the sufficiently small radius ε of the balls $B_0, \ldots, B_d$ and their spatial locations (recall that r_j is drawn from the uniform distribution on B_0 and each b_{i,π_i} is drawn from the uniform distribution on B_i), the point on h closest to r_j (say o) is always inside $\mathcal{CH}(\{b_{1,\pi_1} \ldots b_{d,\pi_d}\})$, no matter what the exact locations of $r_j, b_{1,\pi_1}, \ldots, b_{d,\pi_d}$ are. Thus, o is also the point in $\mathcal{CH}(\{b_{1,\pi_1} \ldots b_{d,\pi_d}\})$ closest to r_j, and the maximum-margin separator of $\{r_j, b_{1,\pi_1} \ldots b_{d,\pi_d}\}$ is the bisector of the segment $[r_j, o]$ by Lemma 8. It then follows that $M_\tau = \mathrm{dist}(r_j, h)/2$. Now let $\tau' = (j', \pi'_1, \ldots, \pi'_d)$ be a tuple different from τ (where $j' \in \{1, \ldots, n\}$ and $\pi'_1, \ldots, \pi'_d \in \{1, \ldots, N/d\}$). Let h' denote the (probabilistic) hyperplane going through $b_{1,\pi'_1}, \ldots, b_{d,\pi'_d}$. We have $M_{\tau'} = \mathrm{dist}(r_{j'}, h')/2$ as argued above. Therefore, to show $\Pr[M_\tau = M_{\tau'}] = 0$, it suffices to show $\Pr[\mathrm{dist}(r_j, h) = \mathrm{dist}(r_{j'}, h')] = 0$. We consider two cases separately: $j \neq j'$ or $j = j'$.

Assume $j \neq j'$. Without loss of generality, we may assume $j = 1$ and $j' = 2$. The idea is to fix the locations of $r_2, \ldots, r_n$ and all blue points, and then show the conditional probability for $\mathrm{dist}(r_j, h) = \mathrm{dist}(r_{j'}, h')$ is 0. Formally, let $r_i^\sim \in B_0$ for $i \in \{2, \ldots, n\}$ and $b_{i,\pi}^\sim \in B_i$ for $i \in \{1, \ldots, d\}$ and $\pi \in \{1, \ldots, N/d\}$ be arbitrary points. We show that

$$\Pr\left[\mathrm{dist}(r_1, h) = \mathrm{dist}(r_2, h') \,\middle|\, \left(\bigwedge_{i=2}^{n} (r_i = r_i^\sim)\right) \wedge \left(\bigwedge_{i=1}^{d} \bigwedge_{\pi=1}^{N/d} (b_{i,\pi} = b_{i,\pi}^\sim)\right)\right] = 0.$$

Since the points $r_i^\sim$'s and $b_{i,\pi}^\sim$'s are arbitrarily chosen, the above equation immediately implies $\Pr[\mathrm{dist}(r_1, h) = \mathrm{dist}(r_2, h')] = 0$. We use Γ to denote the condition in the above conditional probability. Let $h^\sim$ and $h'^\sim$ be the hyperplanes going through $b_{1,\pi_1}^\sim, \ldots, b_{d,\pi_d}^\sim$ and $b_{1,\pi'_1}^\sim, \ldots, b_{d,\pi'_d}^\sim$, respectively. Set $\delta = \mathrm{dist}(r_2^\sim, h'^\sim)$. Then under the condition Γ, $\mathrm{dist}(r_1, h) = \mathrm{dist}(r_2, h')$ iff $\mathrm{dist}(r_1, h^\sim) = \delta$. Since r_1 is uniformly drawn from a ball, it is clear that $\mathrm{dist}(r_1, h^\sim) = \delta$ happens with probability 0. Therefore, $\Pr[\mathrm{dist}(r_1, h) = \mathrm{dist}(r_2, h') | \Gamma] = 0$ and $\Pr[\mathrm{dist}(r_1, h) = \mathrm{dist}(r_2, h')] = 0$.

The other case $j = j'$ is handled similarly by using conditional probability. If $j = j'$, then $\pi_i \neq \pi'_i$ for some $i \in \{1, \ldots, d\}$, since $\tau \neq \tau'$. Without loss of

generality, assume $\pi_1 \neq \pi'_1$. Again, we fix the locations of all random points but b_{1,π_1}, and consider the conditional probability for $\text{dist}(r_j, h) = \text{dist}(r_{j'}, h')$. As in the last paragraph, let $r_i^{\sim} \in B_0$ and $b_{i,\pi}^{\sim} \in B_i$ be arbitrary points. Define Γ as the event that all red points r_i have the locations $r_i^{\sim}$ and all blue points $b_{i,\pi}$ except b_{1,π_1} have the locations $b_{i,\pi}^{\sim}$. We show that $\Pr[\text{dist}(r_j, h) = \text{dist}(r_{j'}, h')|\Gamma] = 0$. Let $h'^{\sim}$ be the hyperplane going through $b_{1,\pi'_1}^{\sim}, \ldots, b_{d,\pi'_d}^{\sim}$, and $\delta = \text{dist}(r_{j'}^{\sim}, h'^{\sim})$. Then under the condition Γ, $\text{dist}(r_j, h) = \text{dist}(r_{j'}, h')$ iff $\text{dist}(r_j^{\sim}, h) = \delta$. Also, $\text{dist}(r_j^{\sim}, h) = \delta$ iff h is tangent to the δ-ball centered at $r_j^{\sim}$. Note that under the condition Γ, h is the hyperplane going through $b_{1,\pi_1}, b_{2,\pi_2}^{\sim}, \ldots, b_{d,\pi_d}^{\sim}$ (in which only b_{1,π_1} is a random point). There are at most two hyperplanes which are tangent to the δ-ball centered at $r_j^{\sim}$ and going through $b_{2,\pi_2}^{\sim}, \ldots, b_{d,\pi_d}^{\sim}$, say h_1 and h_2. So h is tangent to the δ-ball centered at $r_j^{\sim}$ iff $h = h_1$ or $h = h_2$. It follows that h is tangent to the δ-ball centered at $r_j^{\sim}$ only if $b_{1,\pi_1} \in h_1 \cup h_2$. Clearly, the probability of $b_{1,\pi_1} \in h_1 \cup h_2$ is 0, since b_{1,π_1} is uniformly drawn from a ball. Therefore, $\Pr[\text{dist}(r_j, h) = \text{dist}(r_{j'}, h')|\Gamma] = 0$. Because the locations $r_i^{\sim}$ and $b_{i,\pi}^{\sim}$ are arbitrarily chosen, we have $\Pr[\text{dist}(r_j, h) = \text{dist}(r_{j'}, h')] = 0$.

Now we see that $\Pr[M_\tau = M_{\tau'}] = 0$ for any two distinct tuples τ and τ'. By the union bound, we then have

$$\Pr[\ M_\tau = M_{\tau'} \text{ for some } \tau \neq \tau'\ ] \leq \binom{nN^d}{2} \cdot 0 = 0.$$

This further implies that

$$\Pr[\ M_\tau \neq M_{\tau'} \text{ for all } \tau \neq \tau'\ ] = 1.$$

Note that if all M_τ are distinct, then $\kappa(\mathcal{S}^*)$ is at least $n(N/d)^d$. So we conclude

$$\Pr\left[\ \kappa(\mathcal{S}^*) \geq n\left(\frac{N}{d}\right)^d\ \right] \geq \Pr[\ M_\tau \neq M_{\tau'} \text{ for all } \tau \neq \tau'\ ] = 1.$$

This completes the proof of the theorem. □

From the above theorem, we can conclude that any algorithm that explicitly considers every possible separation-margin of the bichromatic stochastic dataset requires at least $\Omega(nN^d)$ time to compute the ESM. This in turn implies that our algorithm is optimal among this category of algorithms. To do better, the only hope

is to avoid considering every possible separation-margin explicitly. However, this is fairly difficult (though it may not be impossible) because of the lack of an explicit relationship among distinct separation margins.

2.4 Extension to general geometric objects

In the previous sections, we studied the separability problems for bichromatic stochastic datesets consisting of only points. In fact, the two problems can be naturally generalized to the case of general geometric objects (see Figure 2.6). In this paper, the general geometric objects to be considered include polytopes with constant number of vertices, and/or d-dim closed balls with various radii. We show that, with some effort, our methods can be extended to solve the generalized versions of the SP and ESM problems. Let $\mathcal{S} = (S_R, S_B, \pi)$ be a bichromatic stochastic datesets consisting of general geometric objects; in other words, each element in S_R and S_B is either a polytope or a ball.

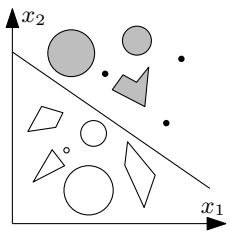


Figure 2.6: A separability problem for a set of bichromatic general objects in $\mathbb{R}^2$

2.4.1 Reducing polytopes to points

To deal with polytopes is easy, because of the fact that the entire polytope is on one side of a (hyperplane) separator iff all its vertices are. Thus, we can simply replace each polytope in S_R and S_B by its vertices and associate with each vertex an existence probability equal to that of the polytope. In this way, we may assume that each element in S_R and S_B is a ball in $\mathbb{R}^d$ (a point can be regarded as a ball of radius 0). It should be noted is that, once we reduce the polytopes to points, the existence of the vertices of each polytope are dependent upon each other, i.e., a realization no longer includes each point independently. However, this issue can be

easily handled without any increase in time complexity, because each polytope only has a constant number of vertices.

2.4.2 Handling balls

Now it suffices to solve the separability problems for the case that $\mathcal{S}$ is a bichromatic stochastic dateset consisting of balls. Before we discuss how to handle balls, we need a definition of general position for a ball-dataset. We say a set of balls in $\mathbb{R}^d$ is in general position (or has the general position property) if **(1)** the centers of the balls are in general position and **(2)** no $(d+1)$ balls have a common tangent hyperplane. Furthermore, we say a ball-dataset has strong general position property (SGPP) if it satisfies the two conditions above and all of the 0-radius balls in it have SGPP (as defined in Section 2.2.1) when regarded as points. When solving the SP problem in Section 2.4.2.1, the given ball-dataset S is required to have SGPP. When solving the ESM problem in Section 2.4.2.2, we only need the assumption that S has the (usual) general position property.

2.4.2.1 Separable-probability (ball-version)

Let $\mathcal{T} = (T_R, T_B)$ be a bichromatic dataset of balls with SGPP and set $J = \{3, 4, \ldots, d\}$. With similar proofs, Lemma 1 and 3 can be directly generalized to ball-datasets (the meaning of $\mathcal{CH}(T_R)$ and $\mathcal{CH}(T_B)$ should be modified as the convex hull of all the balls in T_R and T_B). The ball-version of Lemma 4 (and also its proof) is slightly different, which is presented as follows. We follow the notations used in Section 2.2.1.

Lemma 13. *There exists a unique weak separator for $\mathcal{T}^{U^*}$ in U^*. This separator either goes through exactly d 0-radius balls in $\mathcal{T}^{U^*}$ (of which at least one is in $T_R^{U^*}$ and one is $T_B^{U^*}$) or is tangent to at least one ball in $\mathcal{T}^{U^*}$ of radius larger than 0.*

Proof. By applying the same approach used in the proof of Lemma 4, we can directly show that $\mathcal{T}^{U^*}$ is weakly separable. However, to prove the remaining part, we need to slightly change the approach in the proof of Lemma 4. First, we modify the definition of "degree" in that proof as follows. Let X be the convex hull of a finite set of balls and x be a point on the boundary of X. Also, let Y be the union of

those balls. We define the degree of x in X, denoted by $\deg_X x$, to be the minimum of the dimensions of all the simplices that contain x and use only the points in Y as their vertices. Note that $\deg_X x$ is well-defined. Indeed, as $x \in X = \mathcal{CH}(Y)$, there exists at least one simplex with vertices in Y that contains x. Since $\mathcal{T}^{U^*}$ is not strongly separable, by Lemma 1, there exists a point $x^* \in \mathcal{CH}(T_R^{U^*}) \cap \mathcal{CH}(T_B^{U^*})$. Let $C_1 = \mathcal{CH}(T_R^{U^*})$ and $C_2 = \mathcal{CH}(T_B^{U^*})$. Define $k_1 \deg_{C_1} x^*$ and $k_2 = \deg_{C_2} x^*$. Then we can find a k_1-dim (resp. k_2-dim) simplex $\bar{s}_R$ (resp. $\bar{s}_B$) satisfying
(i) $\bar{s}_R$ (resp. $\bar{s}_B$) contains x^* in its interior;
(ii) each vertex of $\bar{s}_R$ (resp. $\bar{s}_B$) is contained in at least one ball in $T_R^{U^*}$ (resp. $T_B^{U^*}$).
Consider the balls that contain the vertices of $\bar{s}_R$ and $\bar{s}_B$. We have two cases. First, all of those balls are 0-radius balls. Second, at least one of them has the radius larger than 0. In the first case, the proof of Lemma 4 is sufficient to show that the weak separator of $\mathcal{T}^{U^*}$ is unique and goes through d points (0-radius balls). In the second case, without loss of generality, we may assume that there is a vertex v of $\bar{s}_R$ contained in a ball $a \in T_R^{U^*}$ with radius larger than 0. Since any weak separator of $\mathcal{T}^{U^*}$ must go through v, v must be on the boundary of a. Thus, $\mathcal{T}^{U^*}$ has a unique weak separator, which is the tangent hyperplane of a on v (so it is tangent to at least one ball with radius larger than 0). □

Given the above results, we are immediately able to generalize the concept of extreme separator to ball-datasets. As we do in Section 2.2.1, if $P_0 \neq \emptyset$, we define the extreme separator of $\mathcal{T}$ as the derived separator of the unique weak separator of $\mathcal{T}^{U^*}$. If $P_0 = \emptyset$, we say the extreme separator of $\mathcal{T}$ is *not defined*. If the extreme separator is defined, we call the subset of $\mathcal{T}$ consisting of all the balls tangent to extreme separator the *critical set*. Later, we shall use the following lemma to solve the ball version of the SP problem.

Lemma 14. *Let $\mathcal{T} = (T_R, T_B)$ be a separable bichromatic dataset of balls in $\mathbb{R}^d$ whose extreme separator is defined and let $\mathcal{C}$ be its critical set. Then the extreme separator of $\mathcal{C}$ is also defined. Furthermore, $\mathcal{T}$ and $\mathcal{C}$ share the same extreme separator and auxiliary subspace.*

Proof. Recall the ρ-function defined in Section 2.2.1. Let P_0 and P_1 be the pre-images of $\{0\}$ and $\{1\}$ under the map $\rho_{\mathcal{T}}^*$ respectively. Also, let P_0' and P_1' be the

pre-images of $\{0\}$ and $\{1\}$ under the map $\rho^*_{\mathcal{C}}$. Suppose u^* is the clockwise boundary of P_0. Since $\mathcal{C} \subseteq \mathcal{T}$, we have $P'_0 \subseteq P_0$. On the other hand, as $\mathcal{C}$ is the critical set of $\mathcal{T}$, it is easy to see that $\mathcal{CH}(C_R^{U^*}) \cap \mathcal{CH}(C_B^{U^*}) \neq \emptyset$, where $U^* = \sigma(u^*)$. This in turn implies $u^* \in P'_0$. Now because P'_0 is nonempty, the extreme separator of $\mathcal{C}$ is directly defined. Furthermore, from the fact that $u^* \in P'_0 \subseteq P_0$, we know u^* is also the clockwise boundary of P'_0 so that U^* is the auxiliary subspace of both $\mathcal{T}$ and $\mathcal{C}$. To prove $\mathcal{T}$ and $\mathcal{C}$ share the same extreme separator, we assume h is the unique weak separator of $\mathcal{T}^{U^*}$. Since $\mathcal{C}^{U^*} \subseteq \mathcal{T}^{U^*}$, h is also a weak separator of $\mathcal{C}^{U^*}$. More precisely, h is the unique weak separator of $\mathcal{C}^{U^*}$, due to the uniqueness of the weak separator of $\mathcal{C}^{U^*}$ (Lemma 13). Consequently, the derived separator of h in $\mathbb{R}^d$ is the extreme separator of both $\mathcal{T}$ and $\mathcal{C}$. □

Lemma 14 implies that the extreme separator is uniquely determined by the critical set. This then gives us the basic idea to solve the problem: enumerating all possible critical set. As in Section 2.2.2, we can compute the SP of S as

$$SP(S) = SP(\Phi_J(S)) + \sum_{\mathcal{C}} \lambda_{\mathcal{S}}(\mathcal{C}),$$

where $\lambda_{\mathcal{S}}(\mathcal{C})$ is the probability that the critical set of a realization of $\mathcal{S}$ is $\mathcal{C}$. Since the balls in $\mathcal{S}$ have SGPP, the size of the critical set can be at most d. Furthermore, the critical set should contain at least one ball in S_R and one ball in S_B. Thus, it suffices to compute $\lambda_{\mathcal{S}}(\mathcal{C})$ for all the subsets $\mathcal{C} \subseteq (S_R, S_B)$ of size at most d that contains at least one ball in S_R and one ball in S_B. We consider two cases separately. First, all the balls in $\mathcal{C}$ have radius 0. Second, there is at least one ball in $\mathcal{C}$ with radius larger than 0.

In the first case, according to Lemma 13, $\lambda_{\mathcal{S}}(\mathcal{C}) > 0$ only if $\mathcal{C}$ contains exactly d balls. Since the balls in $\mathcal{C}$ are actually points, the situation here is similar to what we confronted in the point-version of the problem. We can uniquely determine a hyperplane h which goes through the d points in $\mathcal{C}$, and a subspace $U^* \in \mathcal{V}$ perpendicular to h. Then $\lambda_{\mathcal{S}}(\mathcal{C})$ is just equal to the probability that h is the extreme separator of the existent balls. The conditions for h to be the extreme separator of a realization R of $\mathcal{S}$ are very similar to those in Section 2.2.2, which are

(i) R contains all the balls in $\mathcal{C}$;

(ii) there exist $r \in \mathcal{CH}(C_R)$ and $b \in \mathcal{CH}(C_B)$ such that their projection images on

U^* coincide;

(iii) R contains no ball in S_R (resp., S_B) that is on the opposite (resp. same) side of h with respect to the point o, where the definition of o is similar to that in Section 2.2.2;

(iv) R contains no ball intersecting with h, except the ones in $\mathcal{C}$.

If $\mathcal{C}$ violates the second condition, then $\lambda_{\mathcal{S}}(\mathcal{C}) = 0$. Otherwise, $\lambda_{\mathcal{S}}(\mathcal{C})$ is just equal to the product of the existence probabilities of the balls in $\mathcal{C}$ and the non-existence probabilities of the balls that R should not contain.

In the second case, however, the size of $\mathcal{C}$ may be less than d. According to Lemma 14, if $\mathcal{C}$ is the critical set of a realization of $\mathcal{S}$, then the extreme separator and auxiliary subspace of the realization are the same as those of $\mathcal{C}$. In particular, this implies that $\lambda_{\mathcal{S}}(\mathcal{C}) = 0$ if $\mathcal{C}$ is not separable or the extreme separator of $\mathcal{C}$ is not defined. So we only need to consider the situation that the extreme separator of $\mathcal{C}$ is defined. Assume that $\mathcal{C}$ has the extreme separator h with the auxiliary subspace $U^* \in \mathcal{V}$. Let c be any ball in $\mathcal{C}$ with radius larger than 0. Then it is easy to see that $\mathcal{C}$ is the critical set of a realization R iff

(i) R contains all the balls in $\mathcal{C}$;

(ii) all the balls in $\mathcal{C}$ are tangent to h;

(iii) R contains no ball with the same color as (resp. different color from) c but on the opposite (resp. same) side of h^* with respect to c;

(iv) R contains no ball intersecting with h, except the ones in $\mathcal{C}$.

Because of the constant size of $\mathcal{C}$, h and U^* can be computed in constant time. Similarly, if $\mathcal{C}$ satisfies the second condition, $\lambda_{\mathcal{S}}(\mathcal{C})$ is equal to the product of the existence probabilities of the balls in $\mathcal{C}$ and the non-existence probabilities of the balls that R should not contain.

In both the cases, $\lambda_{\mathcal{S}}(\mathcal{C})$ can be computed in linear time by simply scanning all the balls in $\mathcal{S}$. Thus, $SP(\mathcal{S})$ can be finally computed in $O(nN^d)$ time, as the number of the subsets $\mathcal{C}$ considered is bounded by $O(nN^{d-1})$. Unfortunately, the improvement techniques used in the point-version of the problem cannot be generalized to ball-datasets so that our eventual time bound for computing the separable-probability of general stochastic geometric objects remains $O(nN^d)$.

Theorem 15. *One can compute the separable-probability of a bichromatic stochastic*

dataset consisting of general geometric objects in $\mathbb{R}^d$ of size (n, N) in $O(nN^d)$ time.

2.4.2.2 Expected separation-margin (ball-version)

Let $\mathcal{T} = (T_R, T_B)$ be a bichromatic dataset of balls in general position. Clearly, the definitions given in Section 2.3 (maximum-margin separator, separation-margin, support set/points/planes, etc.) can be directly generalized to the ball case. Also, with these definitions, the ball-versions of Lemma 8 and 9 can be easily verified (using the same proofs).

To extend the previous algorithm to the ball case, we need to establish the ball version of Lemma 10. The first step is the same as that in the original proof of Lemma 10: we arbitrarily label the balls in $\mathcal{S}$ and define the representation of $\mathcal{C}$ as the $d+1$ balls in $\mathcal{C}$ with the smallest labels, for a subset $\mathcal{C} \subseteq (S_R, S_B)$ of size at least $d+1$. We show that the number of possible support sets represented by any group of $d+1$ balls is $O(1)$. Let $a_1, a_2, \ldots, a_{d+1}$ be any $d+1$ balls in $\mathcal{S}$ where $a_1, \ldots, a_k \in S_R$ and $a_{k+1}, \ldots, a_{d+1} \in S_B$, for some $1 \leq k \leq d$ as before. Suppose each ball a_i has center c_i and radius δ_i. If some possible support set $\mathcal{C}$ is represented by these $d+1$ balls, then the support plane h_r (resp. h_b) must be tangent to $a_1, \ldots, a_k$ (resp. $a_{k+1}, \ldots, a_{d+1}$). Furthermore, the balls $a_1, \ldots, a_k$ (resp. $a_{k+1}, \ldots, a_{d+1}$) must be on the open side of h_r (resp. h_b), i.e., the side different from the one containing the area in between h_r and h_b. Formally, suppose the equations of h_r and h_b are $\vec{\omega} \cdot x + b_1 = 0$ and $\vec{\omega} \cdot x + b_2 = 0$. We then have the following system of equations

$$\begin{cases} \vec{\omega} \cdot c_i + b_1 = -r_i & \text{for } i \in \{1, \ldots, k\}, \\ \vec{\omega} \cdot c_i + b_2 = r_i & \text{for } i \in \{k+1, \ldots, d+1\}, \\ |\vec{\omega}| = 1, \\ b_1 < b_2. \end{cases}$$

The $d+1$ linear equations are linearly independent, as the centers are in general position. Thus, by limiting the norm of $\vec{\omega}$ to be 1, this system has at most two solutions. In other words, there are at most two possibilities for the support planes h_r and h_b. By following the reasoning in the proof of Lemma 10, we then know the number of the possible support sets represented by these $d+1$ balls is constant, which immediately implies that the total number of all possible support sets is bounded by $O(nN^d)$.

To enumerate these possible support sets, we can directly use the same method as in Section 2.3.1, i.e., first enumerate $d+1$ balls and then enumerate the possible support sets represented by them. Again, because the improvement techniques used in the point-version of the problem do not work for ball-datasets, we have to scan all the balls once for computing the corresponding probability of each possible support set, which makes the overall time $O(nN^{d+1})$ for computing the ESM of general geometric objects.

Theorem 16. *One can compute the expected separation-margin of a bichromatic stochastic dataset consisting of general geometric objects in* $\mathbb{R}^d$ *of size* (n, N) *in* $O(nN^{d+1})$ *time.*

Chapter 3

Stochastic convex hull problems

Let $\mathcal{S} = (S, \pi)$ be a given stochastic dataset in $\mathbb{R}^d$ where $S = \{a_1, \ldots, a_n\}$. In this chapter, we study the problems of computing the expected diameter, width, and combinatorial complexity of a stochastic convex hull of $\mathcal{S}$; see Section 1.1 for the statement of these problems.

3.1 Preliminaries

Let P be a convex polytope in $\mathbb{R}^d$. If $\mathbf{u}$ is a unit vector in $\mathbb{R}^d$, we define the *directional width* of P with respect to $\mathbf{u}$ as

$$\mathrm{wid}_{\mathbf{u}}(P) = \sup_{p,q \in P} \left(\langle \mathbf{u}, p \rangle - \langle \mathbf{u}, q \rangle \right),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Let U be the set of unit vectors in $\mathbb{R}^d$. Then the *diameter* of P is defined as $\mathrm{diam}(P) = \sup_{\mathbf{u} \in U} \mathrm{wid}_{\mathbf{u}}(P)$, and the *width* of P is defined as $\mathrm{wid}(P) = \inf_{\mathbf{u} \in U} \mathrm{wid}_{\mathbf{u}}(P)$. It is clear that the diameter of P is also the distance between the farthest-pair of points in P. The *combinatorial complexity* (or simply *complexity*) of P, denoted by $|P|$, is defined as the total number of faces of P (the dimensions of the faces vary from 0 to $d-1$).

For two points $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$ in $\mathbb{R}^d$, we define $x \prec y$ if the d-tuple $(x_1, \ldots, x_d)$ is smaller than the d-tuple $(y_1, \ldots, y_d)$ in lexicographic order. Then $\prec$ induces a (strict) total order on $\mathbb{R}^d$, called $\prec$-*order*.

3.2 Approximating the expected diameter

Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset in $\mathbb{R}^d$ (d is not assumed to be fixed), and suppose $|S| = n$. Our goal in this section is to (approximately) compute the expected diameter of a SCH of $\mathcal{S}$, defined as

$$\text{diam}_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot \text{diam}(\mathcal{CH}(R)),$$

where $\Pr(R)$ denotes the probability that R occurs as a realization of $\mathcal{S}$.

3.2.1 The witness sequence

In order to approximate $\text{diam}_{\mathcal{S}}$, we introduce the notion of *witness sequence*. Let P be a convex polytope in $\mathbb{R}^d$, and V be the vertex set of P. For any point $x \in \mathbb{R}^d$, we define $\Phi_P(x)$ as the set of all points in P farthest from x. Formally, $\Phi_P(x) = \{y \in P : \text{dist}(x, y) \geq \text{dist}(x, y') \text{ for any } y' \in P\}$. Note that $\Phi_P(x) \subseteq V$, and in particular $\Phi_P(x)$ is finite. Our first observation about $\text{diam}(P)$ is the following.

Lemma 17. *Let $x \in \mathbb{R}^d$ be a point. If there exist $p, q \in P$ such that $\text{dist}(p, q) = \text{diam}(P)$ and $\angle pxq = \theta > \pi/2$, then for any $y \in \Phi_P(x)$ and $z \in \Phi_P(y)$ we have*

$$\text{dist}(y, z) \geq \frac{\text{diam}(P)}{2\sin(\pi/2 - \theta/4)}.$$

Proof. Let $x \in \mathbb{R}^d$ be a point, and suppose we have $p, q \in P$ such that $\text{dist}(p, q) = \text{diam}(P)$ and $\angle pxq > \pi/2$. Also, let $y \in \Phi_P(x)$ be any point. Since $\text{dist}(y, z) \geq \max\{\text{dist}(y, p), \text{dist}(y, q)\}$ for any $z \in \Phi_P(y)$, it suffices to show

$$\max\{\text{dist}(y, p), \text{dist}(y, q)\} \geq \frac{\text{diam}(P)}{2\sin(\pi/2 - \theta/4)}.$$

Without loss of generality, we may assume $x = (0, \ldots, 0)$, $p = (\alpha, \beta, 0, \ldots, 0)$, $q = (\alpha, \gamma, 0, \ldots, 0)$, where $\alpha \geq 0$ (if this is not the case, one can properly apply an isometric transformation on $\mathbb{R}^d$ to make it true). Furthermore, we may also assume $\text{dist}(x, y) = 1$, hence $\alpha^2 + \beta^2 \leq 1$ and $\alpha^2 + \gamma^2 \leq 1$. Since $\angle pxq > \pi/2$, we must have $\beta\gamma < 0$ (so suppose $\beta > 0$ and $\gamma < 0$). We first claim that $\max\{\text{dist}(y, p), \text{dist}(y, q)\}$ is minimized when

$$y = \left(\sqrt{1 - \frac{(\beta + \gamma)^2}{4}}, \frac{\beta + \gamma}{2}, 0, \ldots, 0\right). \tag{3.1}$$

Let y be the point with the above coordinates (see Figure 3.1), and $r = (r_1, \ldots, r_d)$ be another point satisfying $\text{dist}(x, r) = 1$ (i.e., $\sum_{i=1}^{d} r_i^2 = 1$). First consider the case of $r_2 \leq (\beta + \gamma)/2$. In this case, we show that $\text{dist}(r, p) \geq \max\{\text{dist}(y, p), \text{dist}(y, q)\}$. Since $\text{dist}(y, p) = \text{dist}(y, q)$, it suffices to show $\text{dist}(r, p) \geq \text{dist}(y, p)$. We have the equations

$$\text{dist}^2(r, p) = 1 + \alpha^2 + \beta^2 - 2r_1\alpha - 2r_2\beta,$$

$$\text{dist}^2(y, p) = 1 + \alpha^2 + \beta^2 - 2y_1\alpha - 2y_2\beta,$$

where y_1 and y_2 are the first two coordinates of y defined above. Now we only need to show $r_1\alpha + r_2\beta \leq y_1\alpha + y_2\beta$. Note that $r_1\alpha + r_2\beta \leq \alpha\sqrt{1 - r_2^2} + r_2\beta$ as $\alpha \geq 0$. Define vectors $\mathbf{v} = (\alpha, \beta)$, $\mathbf{u} = (\sqrt{1 - r_2^2}, r_2)$, $\mathbf{w} = (y_1, y_2)$. Since $\alpha \geq 0$, $y_1 > 0$, and $r_2 \leq y_2 < \beta$, the angle between $\mathbf{v}$ and $\mathbf{u}$ is greater than that between $\mathbf{v}$ and $\mathbf{w}$. Furthermore, $\|\mathbf{u}\|_2 = \|\mathbf{w}\|_2 = 1$. Therefore, $\alpha\sqrt{1 - r_2^2} + r_2\beta = \langle \mathbf{u}, \mathbf{v} \rangle \leq \langle \mathbf{w}, \mathbf{v} \rangle = y_1\alpha + y_2\beta$, which implies $r_1\alpha + r_2\beta \leq y_1\alpha + y_2\beta$. In the other case, $r_2 \geq (\beta + \gamma)/2$, we have symmetrically $\text{dist}(r, q) \geq \max\{\text{dist}(y, p), \text{dist}(y, q)\}$. Therefore, we know that $\max\{\text{dist}(y, p), \text{dist}(y, q)\}$ is minimized when y has the coordinates in Equation 3.1. Note that when y has these coordinates,

$$\text{dist}(y, p) = \text{dist}(y, q) = \frac{\text{dist}(p, q)}{2\sin(\angle pyq/2)} = \frac{\text{diam}(P)}{2\sin(\angle pyq/2)}. \tag{3.2}$$

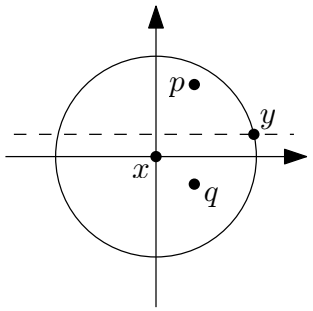


Figure 3.1: The locations of x, p, q and y in the proof of Lemma 17.

Next, we show that $\angle pyq \leq \pi - \theta/2$ where $\theta = \angle pxq$. Since $\text{dist}(x, p) \leq \text{dist}(x, y)$, $\angle xyp \leq \angle xpy$. Also, since $\text{dist}(x, q) \leq \text{dist}(x, y)$, $\angle xyq \leq \angle xqy$. It follows that $\angle pyq = \angle xyp + \angle xyq \leq \angle xpy + \angle xqy$. But $\angle pxq + \angle pyq + \angle xpy + \angle xqy = 2\pi$ and $\angle pxq = \theta$, which implies that $2\angle pyq \leq 2\pi - \theta$, as desired. Using Equation 3.2,

we can conclude that $\mathrm{dist}(y,p) \geq \mathrm{diam}(P)/(2\sin(\pi/2-\theta/4))$, which completes the proof. □

Basically, Lemma 17 states that for a point $x \in \mathbb{R}^d$, if we take $y \in P$ farthest from x and $z \in P$ farthest from y, then the distance between y and z gives us a good approximation for $\mathrm{diam}(P)$ as long as there exists a pair $p, q \in P$ defining $\mathrm{diam}(P)$ with a large angle $\angle pxq$. However, without the existence of such a pair $p, q \in P$, the approximation fails. To handle this, we need our second observation.

Lemma 18. *Let $v \in V$ be a vertex of P, and $u \in \Phi_P(v), w \in \Phi_P(u)$ be two points. Suppose r is the ray with initial point u which goes through v, and x is the point on r which has distance $\mathrm{dist}(u,w)/2$ from u. Then if there exist $p, q \in P$ with $\mathrm{dist}(p,q) = \mathrm{diam}(P)$ and $\angle pxq = \theta$, we have*

$$\mathrm{dist}(u,w) \geq \min\left\{\mathrm{diam}(P), \frac{\mathrm{diam}(P)}{\sqrt{3}\sin(\theta/2)}\right\}.$$

Proof. Let B_v be the (closed) ball centered at u with radius $\mathrm{dist}(v,u)$, and B_u be the (closed) ball centered at u with radius $\mathrm{dist}(u,w)$. Then we have $P \subseteq B_u \cap B_v$, because $u \in \Phi_P(v)$ and $w \in \Phi_P(u)$. Now let r and x be the ray and the point defined in the lemma. Define v' as the point on r which has distance $\mathrm{dist}(u,w)$ from u, so x is the midpoint of the segment connecting v' and u. Set $B_{v'}$ to be the (closed) ball centered at v' with radius $\mathrm{dist}(u,w)$. See Figure 3.2 for an illustration of the balls $B_u, B_v, B_{v'}$. Note that $B_v \subseteq B_{v'}$, since $\mathrm{rad}(B_{v'}) \geq \mathrm{rad}(B_v) + \mathrm{dist}(v,v')$ where $\mathrm{rad}(\cdot)$ denotes the radius of a ball. Therefore, $P \subseteq B_u \cap B_{v'}$. Next, we claim that $B_u \cap B_{v'} \subseteq B_x$, where B_x is the (closed) ball centered at x with radius $\sqrt{3}\cdot\mathrm{dist}(u,w)/2$. Suppose $y \in B_u \cap B_{v'}$ is a point, and assume $\mathrm{dist}(y,u) \geq \mathrm{dist}(y,v')$ without loss of generality (so $\angle yxu \geq \pi/2$). Define $\mu = \mathrm{dist}(u,x)$ and $\gamma = \mathrm{dist}(y,x)$. Then $\gamma = \mu \cdot \sin\angle yux / \sin\angle uyx$. Note that we have the restrictions $\angle yxu \geq \pi/2$ and $\mathrm{dist}(u,y) \leq \mathrm{dist}(u,v') = 2\mu$. Under these restrictions, it is easy to see that γ is maximized when $\mathrm{dist}(u,y) = 2\mu$ and $\angle yxu = \pi/2$. In this case, $\gamma = \sqrt{3}\mu = \mathrm{rad}(B_x)$. Consequently, $B_u \cap B_{v'} \subseteq B_x$, which in turn implies $P \subseteq B_x$. With this observation, we now show the inequality in the lemma. Let $p, q \in P \subseteq B_x$ be two points satisfying $\mathrm{dist}(p,q) = \mathrm{diam}(P)$ and $\angle pxq = \theta$. If $\mathrm{dist}(p,q) \leq \mathrm{dist}(u,w)$, we are done, so assume $\mathrm{dist}(p,q) > \mathrm{dist}(u,w)$. But both $\mathrm{dist}(x,p)$ and $\mathrm{dist}(x,q)$ are at most $\mathrm{rad}(B_x) = \sqrt{3}\cdot\mathrm{dist}(u,w)/2$. Therefore, θ is the largest angle of the

triangle $\triangle pxy$. In this case, it is easy to see that $\mathrm{dist}(p, q)$ is maximized when $\mathrm{dist}(x, p) = \mathrm{dist}(x, q) = \mathrm{rad}(B_x)$. It follows that $\mathrm{dist}(p, q) \leq \sqrt{3}\sin(\theta/2) \cdot \mathrm{dist}(u, w)$, which completes the proof. □

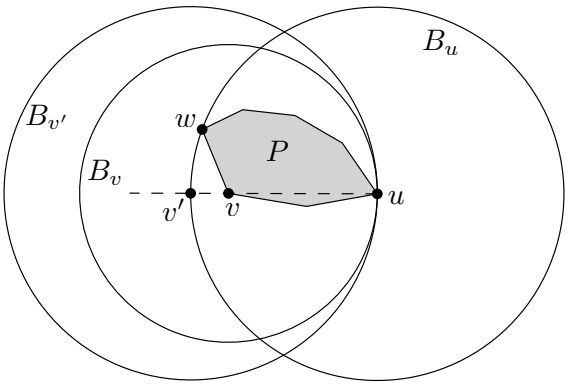


Figure 3.2: An illustration of $B_u, B_v, B_{v'}$ in the proof of Lemma 18.

Lemma 18 states that for a vertex $v \in V$, if we take $u \in P$ farthest from v and $w \in P$ farthest from v, then the distance between u and w gives us a good approximation for $\mathrm{diam}(P)$ as long as there exists a pair $p, q \in P$ defining $\mathrm{diam}(P)$ with a small angle $\angle pxq$ (see the lemma for the definition of x). The approximation is not satisfactory when $\angle pxq$ is large. Fortunately, we already have Lemma 17, which is helpful for this case. Indeed, in the case that $\angle pxq$ is large, if we further take $y \in P$ farthest from x and $z \in P$ farthest from y, then Lemma 17 implies that the distance between y and z is a good approximation for $\mathrm{diam}(P)$. Therefore, intuitively, by taking $\max\{\mathrm{dist}(u, w), \mathrm{dist}(y, z)\}$, we can well-approximate $\mathrm{diam}(P)$ no matter whether $\angle pxq$ is small or large. We formally state this as follows.

Corollary 19. *Let v, u, w, x be the points defined in Lemma 18. Also, let $y \in \Phi_P(x)$ and $z \in \Phi_P(y)$ be any two points. Then we have*

$$\frac{\mathrm{diam}(P)}{2\sqrt{2}/\sqrt{3}} \leq \max\{\mathrm{dist}(u, w), \mathrm{dist}(y, z)\} \leq \mathrm{diam}(P).$$

Proof. It is clear that $\max\{\mathrm{dist}(u, w), \mathrm{dist}(y, z)\} \leq \mathrm{diam}(P)$, because $u, w, y, z \in P$. Let $p, q \in P$ be two points such that $\mathrm{dist}(p, q) = \mathrm{diam}(P)$. Set $\theta = \angle pxq$. If $\theta \leq \pi/2$, then Lemma 18 implies $\mathrm{dist}(u, w) \geq \mathrm{diam}(P)/(\sqrt{3}/\sqrt{2})$. So assume $\theta > \pi/2$. By

Lemma 17, we have

$$\mathrm{dist}(y,z) \geq \frac{\mathrm{diam}(P)}{2\sin(\pi/2-\theta/4)} = \frac{\mathrm{diam}(P)}{2\cos(\theta/4)}.$$

Also, by Lemma 18, we have $\mathrm{dist}(u,w) \geq \mathrm{diam}(P)/(\sqrt{3}\sin(\theta/2))$. Therefore,

$$\max\{\mathrm{dist}(u,w), \mathrm{dist}(y,z)\} \geq \frac{\mathrm{diam}(P)}{\min\{2\cos(\theta/4), \sqrt{3}\sin(\theta/2)\}}.$$

Note that for $\theta \in (\pi/2, \pi]$, the function $2\cos(\theta/4)$ is monotonically decreasing and the function $\sqrt{3}\sin(\theta/2)$ is monotonically increasing. Thus, the right side of the above inequality is minimized when $2\cos(\theta/4) = \sqrt{3}\sin(\theta/2)$. We have this equality when $\sin(\theta/4) = 1/\sqrt{3}$, because $\sin(\theta/2) = 2\sin(\theta/4)\cos(\theta/4)$. By some direct calculations, we obtain the inequality in the corollary. □

With the five points v, u, w, y, z (which are in fact the vertices of P) in hand, Corollary 19 allows us to approximate $\mathrm{diam}(P)$ within a factor of $2\sqrt{2}/\sqrt{3} \approx 1.633$. In other words, the diameter information of P is well "encoded" in those five vertices. However, the choice of v, u, w, y, z is not unique in our above construction. For later use, we need to make it unique, which can be easily done by considering $\prec$-order. We define $v \in V$ as the largest vertex of P under $\prec$-order. Also, we require $u \in \Phi_P(v)$, $w \in \Phi_P(u)$, $y \in \Phi_P(x)$, $z \in \Phi_P(y)$ to be the largest under $\prec$-order. In this way, we obtain a uniquely defined 5-tuple (v, u, w, y, z) for the polytope P. We call this 5-tuple the *witness sequence* of P, denoted by ψ_P. For a 5-tuple $\psi = (x_1, \ldots, x_5)$ of points in $\mathbb{R}^d$, define $\Lambda(\psi) = \max\{\mathrm{dist}(x_2, x_3), \mathrm{dist}(x_4, x_5)\}$. Then Corollary 19 implies

$$\frac{\mathrm{diam}(P)}{2\sqrt{2}/\sqrt{3}} \leq \Lambda(\psi_P) \leq \mathrm{diam}(P) \tag{3.3}$$

for any convex polytope P in $\mathbb{R}^d$.

3.2.2 An (n, d)-polynomial-time approximation algorithm

With the notion of witness sequence defined above, we can now present our algorithm for the expected-diameter problem. Given the stochastic dataset $\mathcal{S} = (S, \pi)$, we first do a preprocessing to sort all the points in S in $\prec$-order and compute the pair-wise distances of the points in S. This preprocessing can be done in $O(dn^2)$ time. Now

we consider how to approximate $\mathrm{diam}_{\mathcal{S}}$. We define

$$\mathrm{diam}^*_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot \Lambda(\psi_{\mathcal{CH}(R)}).$$

Inequality 3.3 implies $\mathrm{diam}_{\mathcal{S}}/(2\sqrt{2}/\sqrt{3}) \leq \mathrm{diam}^*_{\mathcal{S}} \leq \mathrm{diam}_{\mathcal{S}}$. Thus, in order to achieve a 1.633-approximation $\mathrm{diam}_{\mathcal{S}}$, it suffices to compute $\mathrm{diam}^*_{\mathcal{S}}$. Computing $\mathrm{diam}^*_{\mathcal{S}}$ by directly using the above formula takes exponential time, as S has 2^n subsets. However, since for any $R \subseteq S$ the witness sequence $\psi_{\mathcal{CH}(R)}$ must be a 5-tuple of points in S, we can also write $\mathrm{diam}^*_{\mathcal{S}}$ as

$$\mathrm{diam}^*_{\mathcal{S}} = \sum_{\psi \in \Psi_S} \Pr(\psi) \cdot \Lambda(\psi), \tag{3.4}$$

where Ψ_S is the set of all 5-tuples of points in S and

$$\Pr(\psi) = \Pr_{R \sim \mathcal{S}}[\psi_{\mathcal{CH}(R)} = \psi]$$

is the probability that the witness sequence of a SCH of $\mathcal{S}$ is ψ. Note that $|\Psi_S| = O(n^5)$. Thus, we can efficiently compute $\mathrm{diam}^*_{\mathcal{S}}$ as long as $\Pr(\psi)$ and $\Lambda(\psi)$ can be computed efficiently for every $\psi \in \Psi_S$. Clearly, $\Lambda(\psi)$ can be directly computed in constant time (after our preprocessing). To compute $\Pr(\psi)$, suppose $\psi = (p_1, \ldots, p_5) \in \Psi_S$. It is easy to check that if $p_1 = p_2$, then either $\Pr(\psi) = 0$ or $\Lambda(\psi) = 0$. So we may assume $p_1 \neq p_2$. In this case, we give the following criterion for checking if ψ is the witness sequence of a SCH of $\mathcal{S}$. For three points $a, b, c \in \mathbb{R}^d$, we write $a \prec_b c$ if $\mathrm{dist}(a, b) < \mathrm{dist}(c, b)$, or $\mathrm{dist}(a, b) = \mathrm{dist}(c, b)$ and $a \prec c$.

Lemma 20. *Let $\psi = (p_1, \ldots, p_5) \in \Psi_S$ with $p_1 \neq p_2$. Suppose r is the ray with initial point p_2 which goes through p_1, and x is the point on r which has distance $\mathrm{dist}(p_2, p_3)/2$ from p_2. For a realization R of $\mathcal{S}$, we have $\psi = \mathrm{wit}(\mathcal{CH}(R))$ iff the following two conditions hold.*
(1) *R contains $p_1, \ldots, p_5$.*
(2) *R does not contain any point $a \in S$ satisfying $p_1 \prec a$ or $p_2 \prec_{p_1} a$ or $p_3 \prec_{p_2} a$ or $p_4 \prec_x a$ or $p_5 \prec_{p_4} a$.*

Proof. Let R be a realization of $\mathcal{S}$, and set $C = \mathcal{CH}(R)$. The proof of the lemma is somewhat straightforward by using the definition of witness sequence. To see the "if" part, assume the two conditions in the lemma hold. Then p_1 must be

the largest point in R under $\prec$-order, which must be a vertex of C. Furthermore, p_2, p_3, p_4, p_5 must be the largest points in $\Phi_C(p_1), \Phi_C(p_2), \Phi_C(x), \Phi_C(p_4)$ under $\prec$-order, respectively. Thus, by definition, $\psi = (p_1, \ldots, p_5) = \psi_C$. To see the "only if" part, assume $\psi_C = \psi$. Then $p_1, \ldots, p_5$ are vertices of C and must be contained in R, which implies condition (1). By definition, p_1 is the largest vertex of C under $\prec$-order, and p_2, p_3, p_4, p_5 are the largest points in $\Phi_C(p_1), \Phi_C(p_2), \Phi_C(x), \Phi_C(p_4)$ under $\prec$-order respectively, which implies condition (2). □

By Lemma 20, it is quite easy to compute $\Pr(\psi)$ in linear time, just by multiplying the existence probabilities of the points in ψ and the non-existence probabilities of all the points which should not be included in R (according to condition (2) in the lemma). Using Equation 3.4, we obtain an (n, d)-polynomial-time algorithm to compute diam^*_S. This algorithm runs in $O(n^6 + dn^2)$ time. But we can easily improve the runtime to $O(n^5 \log n + dn^2)$ as follows. Fixing $p_1, p_2, p_3, p_4 \in S$, we show how to compute $\Pr(\psi)$ for all $\psi \in \Psi_S$ of the form $\psi = (p_1, \ldots, p_4, \cdot)$ in $O(n \log n)$ time. As argued before, we may assume $p_1 \neq p_2$. Let r be the ray with initial point p_2 which goes through p_1, and x be the point on r which has distance $\mathrm{dist}(p_2, p_3)/2$ from p_2. First, we determine a subset $A \subseteq S$ consisting of p_4 and all the points $a \in S$ satisfying $p_1 \prec a$ or $p_2 \prec_{p_1} a$ or $p_3 \prec_{p_2} a$ or $p_4 \prec_x a$. It is clear that $\Pr(\psi) > 0$ for $\psi = (p_1, \ldots, p_4, q)$ only if $q \in S \backslash A$. For each $q \in S \backslash A$, we denote by B_q the set of all points $b \in S \backslash A$ with $q \prec_{p_4} b$. By Lemma 20, we have

$$\Pr(\psi_q) = \left(\prod_{i=1}^{4} \pi(p_i) \cdot \prod_{a \in A} (1 - \pi(a)) \right) \cdot \left(\pi(q) \cdot \prod_{b \in B_q} (1 - \pi(b)) \right), \qquad (3.5)$$

where $\psi_q = (p_1, p_2, p_3, p_4, q)$. Note that the left part of the above formula is independent of q and thus only needs to be computed once. To compute the right part efficiently, suppose $S \backslash A = \{c_1, \ldots, c_r\}$. We relabel these points such that $c_1 \prec_{p_4} \cdots \prec_{p_4} c_r$. This can be done by sorting in $O(n \log n)$ time, or more precisely, $O(r \log r)$ time. We then compute $\prod_{j=i}^{r} (1 - \pi(c_j))$ for all $i \in \{1, \ldots, r\}$ (note that this can be done in linear time). With this in hand, we consider each $q \in S \backslash A$. We must have $q = c_i$ for some $i \in \{1, \ldots, r\}$. In this case, the right part of Equation 3.5 is just $\pi(c_i) \cdot \prod_{j=i+1}^{r} (1 - \pi(c_j))$ and hence can be computed in constant time. Therefore, we can compute $\Pr(\psi_q)$ for all $q \in S \backslash A$ in linear time. Including the time for

sorting, this gives us the $O(n^5 \log n)$-time 1.633-approximation algorithm for computing $\mathrm{diam}_{\mathcal{S}}$.

Theorem 21. *One can achieve a* 1.633*-approximation of* $\mathrm{diam}_{\mathcal{S}}$ *in* (n, d)*-polynomial time. Specifically, the approximation can be done in* $O(n^5 \log n + dn^2)$ *time.*

3.2.3 A polynomial-time approximation scheme

In this section, we design a PTAS for computing $\mathrm{diam}_{\mathcal{S}}$. We first consider a special case in which $\mathrm{diam}_{\mathcal{S}} = \Omega(\mathrm{diam}(S))$ and then consider the general case. Let c be a constant such that $\mathrm{diam}_{\mathcal{S}} \geq c \cdot \mathrm{diam}(S)$. We first compute an 2-approximation, $\mathrm{diam}^{\sim}(S)$, of $\mathrm{diam}(S)$ in $O(n)$ time. This can be done by taking an arbitrary point $a \in S$ and the point $b \in S$ farthest to a and defining $\mathrm{diam}^{\sim}(S) = \|a - b\|_2$. Then we build a grid Γ on $\mathbb{R}^d$ consisting of hyper-cube cells with side-length $(c\varepsilon/4) \cdot \mathrm{diam}^{\sim}(S)$, where ε is the approximation factor of our PTAS. If $\square$ is a cell of Γ, we define the notation $S_{\square} = S \cap \square$. A cell $\square$ of Γ is called *nonempty* if $S_{\square} \neq \emptyset$. For each nonempty cell $\square$ of Γ, let $c_{\square}$ denote its center. Now we construct a new stochastic dataset (equipped with existential uncertainty) $\mathcal{S}' = (S', \pi')$ as $S' = \{c_{\square} : \square \text{ is a nonempty cell of } \Gamma\}$ and

$$\pi'(c_{\square}) = 1 - \prod_{a \in S_{\square}} (1 - \pi(a)).$$

We have the following observation.

Lemma 22. $(1 - \varepsilon) \cdot \mathrm{diam}_{\mathcal{S}} \leq \mathrm{diam}_{\mathcal{S}'} \leq (1 + \varepsilon) \cdot \mathrm{diam}_{\mathcal{S}}$.

Proof. Consider the map $f : S \to S'$ defined as $f(a) = c_{\square}$ where $\square$ is the cell containing a. Note that for any $a \in S$, we have $\|a - f(a)\|_2 \leq (c\varepsilon/2) \cdot \mathrm{diam}(S)$. A subset P of S is then mapped to a subset $P' = f(P)$ of S'. Suppose the points p, q define $\mathrm{diam}(P)$. Then we have

$$\mathrm{diam}(P') \geq \|f(p) - f(q)\|_2 \geq \|p - q\|_2 - \|p - f(p)\|_2 - \|q - f(q)\|_2 \geq \mathrm{diam}(P) - \varepsilon \cdot \mathrm{diam}(S).$$

Using a similar argument, we can also deduce that $\mathrm{diam}(P) \geq \mathrm{diam}(P') - \varepsilon \cdot \mathrm{diam}(S)$. Therefore, we have $|\mathrm{diam}(P) - \mathrm{diam}(P')| \leq c\varepsilon \cdot \mathrm{diam}(S)$. Furthermore, it is easy to verify the equation

$$\Pr_{R' \sim \mathcal{S}'}[R' = P'] = \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}}[R = P]$$

from the existence-probability function π' defined above. It follows that

$$\mathrm{diam}_{\mathcal{S}'} = \sum_{P' \subseteq S'} \Pr_{R' \sim \mathcal{S}'}[R' = P'] \cdot \mathrm{diam}(P') = \sum_{P' \subseteq S'} \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}}[R = P] \cdot \mathrm{diam}(P').$$

On the other hand, we have

$$\mathrm{diam}_{\mathcal{S}} = \sum_{P \subseteq S} \Pr_{R \sim \mathcal{S}}[R = P] \cdot \mathrm{diam}(P) = \sum_{P' \subseteq S'} \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}}[R = P] \cdot \mathrm{diam}(P).$$

Therefore,

$$|\mathrm{diam}_{\mathcal{S}'} - \mathrm{diam}_{\mathcal{S}}| \leq \sum_{P' \subseteq S'} \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}}[R = P] \cdot |\mathrm{diam}(P') - \mathrm{diam}(P)| \leq c\varepsilon \cdot \mathrm{diam}(S).$$

Since $c\varepsilon \cdot \mathrm{diam}(S) \leq \varepsilon \cdot \mathrm{diam}_{\mathcal{S}}$, we have $(1-\varepsilon) \cdot \mathrm{diam}_{\mathcal{S}} \leq \mathrm{diam}_{\mathcal{S}'} \leq (1+\varepsilon) \cdot \mathrm{diam}_{\mathcal{S}}$. □

Using the above lemma to approximate $\mathrm{diam}_{\mathcal{S}}$, it suffices to compute $\mathrm{diam}_{\mathcal{S}'}$. This is much easier because of the small size of $\mathcal{S}'$. We notice that $|S'| = O(\varepsilon^{-d})$. Indeed, $|S'|$ is just the number of the nonempty cells of Γ, which is bounded by $O(\varepsilon^{-d})$ since the side-length of the cells is $O(\varepsilon \cdot \mathrm{diam}(S))$. Therefore, we can apply brute-force to compute $\mathrm{diam}_{\mathcal{S}'}$ in $O(\varepsilon^{-d} \cdot 2^{\varepsilon^{-d}})$ time. Including the time for constructing $\mathcal{S}'$, the total time cost of the above procedure is $O(n + \varepsilon^{-d} \cdot 2^{\varepsilon^{-d}})$.

Next, we consider the general case. Let R be a realization of $\mathcal{S}$. For $i, j \in \{1, \ldots, n\}$, define $E_{i,j}$ as the event that the point in R with the smallest index is a_i and the point in R farthest from a_i is a_j. Let $\mathcal{E} = \{E_{i,j} : i \neq j\}$. We can write

$$\mathrm{diam}_{\mathcal{S}} = \sum_{E_{i,j} \in \mathcal{E}} \Pr_{R \sim \mathcal{S}}[E_{i,j}] \cdot \mathbb{E}_{R \sim \mathcal{S}}[\mathrm{diam}(R) | E_{i,j}]. \tag{3.6}$$

Note that $E_{i,j}$ happens iff a_i, a_j exist and all the points in $T_{i,j} = \{a_k : k < i \text{ or } \mathrm{dist}(a_k, a_i) > \mathrm{dist}(a_j, a_i)\}$ do not exist, which implies that $\Pr_{R \sim \mathcal{S}}[E_{i,j}]$ is equal to the product of the nonexistence probabilities of the points in $T_{i,j}$. Therefore, $\Pr_{R \sim \mathcal{S}}[E_{i,j}]$ can be computed efficiently in $O(n)$ time. To approximate $\mathrm{diam}_{\mathcal{S}}$, it suffices to (approximately) compute $\mathbb{E}_{R \sim \mathcal{S}}[\mathrm{diam}(R) | E_{i,j}]$ for all $E_{i,j} \in \mathcal{E}$. We show that this can be solved using our previous algorithm for the case $\mathrm{diam}_{\mathcal{S}} = \Omega(\mathrm{diam}(S))$. For each $E_{i,j} \in \mathcal{E}$, we define a stochastic dataset $\mathcal{S}_{i,j} = (S_{i,j}, \pi_{i,j})$ where $S_{i,j} = S \backslash T_{i,j}$ and

$$\pi_{i,j}(a) = \begin{cases} 1 & \text{if } a = a_i \text{ or } a = a_j, \\ \pi(a) & \text{otherwise.} \end{cases}$$

As argued before, $E_{i,j}$ happens iff a_i, a_j exist and all the points in $T_{i,j}$ do not exist. Thus, we have the equation $\mathbb{E}_{R\sim\mathcal{S}}[\text{diam}(R)|E_{i,j}] = \text{diam}_{\mathcal{S}_{i,j}}$. We claim that $\text{diam}_{\mathcal{S}_{i,j}} = \Omega(\text{diam}(S_{i,j}))$, whence we can compute an approximation of $\text{diam}_{\mathcal{S}_{i,j}}$ using our algorithm mentioned above. First, we have $\text{diam}(S_{i,j}) \leq 2\text{dist}(a_i, a_j)$ because a_j is the point in $S_{i,j}$ farthest from a_i and hence all the points in $S_{i,j}$ are contained in the ball centered at a_i with radius $\text{dist}(a_i, a_j)$. On the other hand, we have $\text{diam}_{\mathcal{S}_{i,j}} \geq \text{dist}(a_i, a_j)$ since every possible realization of $\mathcal{S}_{i,j}$ contains a_i and a_j. As a result, $\text{diam}_{\mathcal{S}_{i,j}} = \Omega(\text{diam}(S_{i,j}))$. With this observation in hand, we can apply our previous algorithm to approximate $\text{diam}_{\mathcal{S}_{i,j}}$. Combining this with Equation 3.6, we obtain a PTAS for computing $\text{diam}_{\mathcal{S}}$.

Theorem 23. *There exists a PTAS for computing the expected diameter of a stochastic data set in* $\mathbb{R}^d$.

3.2.4 #P-hardness of the expected-diameter problem

We prove the #P-hardness of computing $\text{diam}_{\mathcal{S}}$ exactly when the dimension d is not assumed to be fixed. Our result strengthens a result in [8] which states that computing the expected farthest-pair distance of a stochastic dataset in a (general) metric space is #P-hard.

By the definition of #P-hardness, it suffices to give a polynomial-time reduction from some known #P-hard problem to the the problem of computing $\text{diam}_{\mathcal{S}}$. Our reduction is from the problem of counting independent sets of a graph, which is a well-known #P-hard problem. We first establish the following two lemmas.

Lemma 24. *For an integer $k > 0$, there exists two positive real numbers α_k, β_k with $\alpha_k < \beta_k$ and a map $f : \{0, 1, \ldots, k, k+1\} \to \mathbb{R}^k$ such that $\text{dist}(f(i), f(j)) = \alpha_k$ for any $i \neq j$ except that $\text{dist}(f(k), f(k+1)) = \text{dist}(f(k+1), f(k)) = \beta_k$.*

Proof. Let Δ be a regular k-simplex (i.e., a k-simplex with edges of length 1) with vertices $v_0, \ldots, v_k$, Δ' be another regular k-simplex with vertices $v'_0, \ldots, v'_k$. We form a *regular double-simplex* by identically gluing the face $(v_0, \ldots, v_{k-1})$ of Δ with the face $(v'_0, \ldots, v'_{k-1})$ of Δ' (see Figure 3.3). Clearly, this double-simplex can be (isometrically) embedded into $\mathbb{R}^k$ via an embedding map σ. Now we define $f(k) = \sigma(v_k)$, $f(k+1) = \sigma(v'_k)$, and $f(i) = \sigma(v_i) = \sigma(v'_i)$ for all $i \in \{0, \ldots, k-1\}$.

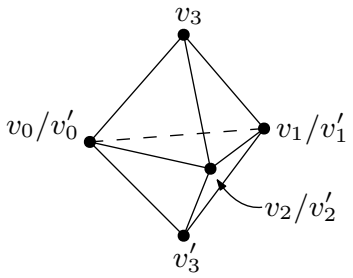


Figure 3.3: The regular double-simplex in the proof of Lemma 24.

By taking $\alpha_k = \mathrm{dist}(f(0), f(1))$ and $\beta_k = \mathrm{dist}(f(k), f(k+1))$, we complete the proof (the desired properties of α_k, β_k, f can be easily verified). □

Lemma 25. *For a graph $G = (V, E)$, one can compute in polynomial time a map $f : V \to \mathbb{R}^{|V|-1}$ such that*

$$\mathrm{dist}(f(u), f(v)) = \begin{cases} \alpha & \text{if } (u,v) \notin E, \\ \beta & \text{if } (u,v) \in E, \end{cases}$$

for some α, β with $\alpha < \beta$.

Proof. Suppose $V = \{v_1, \ldots, v_n\}$ and $E = \{e_1, \ldots, e_m\}$. Using Lemma 24, we find the real numbers $\alpha_{n-2}, \beta_{n-2}$. For each $e \in E$, let $g_e : V \to \mathbb{R}^{n-2}$ be a map such that

$$\mathrm{dist}(g_e(u), g_e(v)) = \begin{cases} \alpha_{n-2} & \text{if } e \neq (u,v), \\ \beta_{n-2} & \text{if } e = (u,v). \end{cases}$$

Note that g_e exists by Lemma 24. We then define $g : V \to (\mathbb{R}^{n-2})^m$ by setting $g(v) = (g_{e_1}(v), \ldots, g_{e_m}(v))$. Let $\alpha = \sqrt{m} \cdot \alpha_{n-2}$ and $\beta = \sqrt{(m-1)\alpha_{n-2} + \beta_{n-2}}$. It is easy to check that $\alpha < \beta$ and

$$\mathrm{dist}(g(u), g(v)) = \begin{cases} \alpha & \text{if } (u,v) \notin E, \\ \beta & \text{if } (u,v) \in E. \end{cases}$$

To further construct f, we note that the image of g consists of only n points, which should span a $(n-1)$-dim hyperplane in $(\mathbb{R}^{n-2})^m$. If we (isometrically) identify this hyperplane with $\mathbb{R}^{n-1}$ and use $h : (\mathbb{R}^{n-2})^m \to \mathbb{R}^{n-1}$ to denote the projection map, $f : V \to \mathbb{R}^{n-1}$ is constructed as the composition $h \circ g$. □

With the above result in hand, we can now describe the reduction. Given a graph $G = (V, E)$ with $V = \{v_1, \ldots, v_n\}$, we first use Lemma 25 to compute the function $f : V \to \mathbb{R}^{n-1}$ and obtain α, β. Let S be the n points in the image of f. We construct a stochastic dataset $\mathcal{S} = (S, \pi)$ by defining $\pi : S \to (0, 1]$ as $\pi(a) = 0.5$ for all $a \in S$. Now the subsets of V are in one-to-one correspondence with the realizations of $\mathcal{S}$. By the construction of f, it is clear that a realization $R \subseteq S$ has a diameter $\mathrm{diam}(R) = \alpha$ if R corresponds to an independent set of G, and has a diameter $\mathrm{diam}(R) = \beta$ otherwise. Furthermore, every subset of S occurs as a realization with an equal probability 2^{-n}. Hence, we immediately obtain the equation

$$\mathrm{diam}_{\mathcal{S}} = \beta + 2^{-n} Ind(G) \cdot (\alpha - \beta),$$

where $Ind(G)$ is the number of the independent sets of G. In this way, counting the independent sets of G is reduced to computing $\mathrm{diam}_{\mathcal{S}}$, which implies the following hardness result.

Theorem 26. *Computing* $\mathrm{diam}_{\mathcal{S}}$ *is #P-hard if the dimension* d *is not fixed.*

Note that our reduction above does not work when d is fixed, as the stochastic dataset $\mathcal{S}$ that we construct is $(n-1)$-dimensional, where n is the number of the points and is not fixed.

3.3 Approximating the expected width

Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset in $\mathbb{R}^d$ with d fixed, and suppose $|S| = n$. Our goal in this section is to (approximately) compute the expected width of a SCH of $\mathcal{S}$, defined as

$$\mathrm{wid}_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot \mathrm{wid}(\mathcal{CH}(R)),$$

where $\Pr(R)$ denotes the probability that R occurs as a realization of $\mathcal{S}$.

3.3.1 The witness simplex

Recall that when solving the expected-diameter problem, we developed the notion of witness sequence, which well-captures the diameter of a polytope and satisfies (1) the total number of the possible witness sequences of a SCH is polynomial (though

there are exponentially many realizations), and (2) the probability of a sequence being the witness sequence of a SCH can be easily computed. We apply this basic idea again to the expected-width problem by defining the so-called *witness simplex*. Let P be a convex polytope in $\mathbb{R}^d$ with $\mathrm{wid}(P) > 0$, and V be the vertex set of P. We choose $d+1$ vertices $v_0, \ldots, v_d \in V$ of P inductively as follows. Define $v_0 \in V$ as the largest vertex of P under $\prec$-order. Suppose $v_0, \ldots, v_i$ are already defined. Let E_i be the (unique) i-dim hyperplane in $\mathbb{R}^d$ through $v_0, \ldots, v_i$ (or the i-dim linear subspace of $\mathbb{R}^d$ spanned by $v_0, \ldots, v_i$). We then define $v_{i+1} \in V$ as the vertex of P which has the maximum distance to E_i, i.e., $v_{i+1} = \arg\max_{v \in V} \mathrm{dist}(v, E_i)$. If there exist multiple vertices having maximum distance to E_i, we choose the largest one under $\prec$-order to be v_{i+1}. In this way, we obtain the vertices $v_0, \ldots, v_d$. The *witness simplex* Δ_P of P is defined as the d-simplex with vertices $v_0, \ldots, v_d$. The (ordered) sequence $(v_0, \ldots, v_d)$ is said to be the *vertex list* of Δ_P. Note that the vertex list is determined by only Δ_P and is independent of P. In other words, if we only know Δ_P without knowing the original polytope P, we can still recover the vertex list of Δ_P, just by ordering the $d+1$ vertices of Δ_P into a sequence $(v_0, \ldots, v_d)$ such that v_0 is the largest under $\prec$-order, and each v_{i+1} is the one having the maximum distance to E_i (the linear subspace spanned by $v_0, \ldots, v_i$). A useful geometric property of the witness simplex Δ_P is that it well-captures the width of P.

Lemma 27. *Let P be a convex polytope in $\mathbb{R}^d$ with $\mathrm{wid}(P) > 0$, then we have $\mathrm{wid}_{\mathbf{u}}(\Delta_P) = \Theta(\mathrm{wid}_{\mathbf{u}}(P))$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$, and in particular, $\mathrm{wid}(\Delta_P) = \Theta(\mathrm{wid}(P))$. (The constant hidden in $\Theta(\cdot)$ could be exponential in d.)*

Proof. Note that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \leq \mathrm{wid}_{\mathbf{u}}(P)$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$, since $\Delta_P \subseteq P$. It suffices to show that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) = \Omega(\mathrm{wid}_{\mathbf{u}}(P))$. Let $(v_0, \ldots, v_d)$ be the vertex list of Δ_P. Also, let E_i be the i-dim hyperplane in $\mathbb{R}^d$ through $v_0, \ldots, v_i$. Suppose each v_i has the coordinates $v_i = (y_{i,1}, \ldots, y_{i,d})$. Without loss of generality, we may assume that $y_{i,j} = 0$ for $j > i$, that is, $v_0 = (0, \ldots, 0)$, $v_1 = (y_{1,1}, 0, \ldots, 0)$, $v_2 = (y_{2,1}, y_{2,2}, 0, \ldots, 0)$, and so forth (if this is not the case, one can properly apply an isometric transformation on $\mathbb{R}^d$ to make it true). With this assumption, E_i is nothing but the i-dim linear subspace of $\mathbb{R}^d$ spanned by the axes $x_1, \ldots, x_i$. Note that $|y_{i,i}| = \mathrm{dist}(v_i, E_{i-1}) \geq \mathrm{dist}(v_{i+1}, E_{i-1}) \geq |y_{i+1,i+1}|$. Therefore, $|y_{1,1}| \geq \cdots \geq |y_{d,d}|$. Furthermore, let $v \in V$ be any vertex of P with coordinates $v = (z_1, \ldots, z_d)$.

For every $i \in \{1, \ldots, d\}$, we have that $\mathrm{dist}(v_i, E_{i-1}) \geq \mathrm{dist}(v, E_{i-1}) \geq |z_i|$, which implies $-|y_{i,i}| \leq z_i \leq |y_{i,i}|$. Based on this observation, we now show that $\mathrm{wid}(\Delta_P) \geq c \cdot \mathrm{wid}(P)$ for some constant c. It suffices to show that there exists a constant c such that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c \cdot \mathrm{wid}_{\mathbf{u}}(P)$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$. We use induction to achieve this. First, for $\mathbf{u} = (0, \ldots, 0, 1)$, we have

$$\mathrm{wid}_{\mathbf{u}}(\Delta_P) = |y_{d,d}| \geq \mathrm{wid}_{\mathbf{u}}(P)/2,$$

because the d-th coordinate of any $v \in V$ has absolute value at most $|y_{d,d}|$. It follows that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_d \cdot \mathrm{wid}_{\mathbf{u}}(P)$ for a constant $c_d = 1/2$. Using this as a base case, we may assume that there exists a constant $c_{i+1} \in (0, 1)$ such that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_{i+1} \cdot \mathrm{wid}_{\mathbf{u}}(P)$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$ whose first i coordinates are all 0. Our goal is to find a new constant $c_i \in (0, 1)$ such that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$ whose first $i-1$ coordinates are all 0. Let $\mathbf{u} = (0, \ldots, 0, u_i, \ldots, u_d) \in \mathbb{R}^d$ be such a unit vector, and define $\mathbf{u}' = (0, \ldots, 0, u'_{i+1}, \ldots, u'_d) \in \mathbb{R}^d$ as a unit vector where $u'_j = u_j / \sqrt{1 - u_i^2}$ for $j \in \{i+1, \ldots, d\}$. We may assume $u_i \geq 0$ because $\mathrm{wid}_{\mathbf{u}}(\Delta_P) = \mathrm{wid}_{-\mathbf{u}}(\Delta_P)$. Set $c_i = c_{i+1}/5$. We verify that $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$ by considering two cases, $u_i|y_{i,i}| \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$ and $u_i|y_{i,i}| < c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$. In the case of $u_i|y_{i,i}| \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$, we immediately have

$$\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq |\langle \mathbf{u}, v_i \rangle - \langle \mathbf{u}, v_{i-1} \rangle| = u_i|y_{i,i}| \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P).$$

In the case of $u_i|y_{i,i}| < c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$, we consider the unit vector $\mathbf{u}'$ defined above. Let $\alpha, \beta \in \{0, \ldots, d\}$ be indices such that $\mathrm{wid}_{\mathbf{u}'}(\Delta_P) = \langle \mathbf{u}', v_\alpha \rangle - \langle \mathbf{u}', v_\beta \rangle$. We claim that $\langle \mathbf{u}, v_\alpha \rangle - \langle \mathbf{u}, v_\beta \rangle \geq c_i \cdot \mathrm{wid}_{\mathbf{u}}(P)$. First, since the i-th coordinates of v_α and v_β have absolute values at most $|y_{i,i}|$ (as observed before), we have

$$\langle \mathbf{u}, v_\alpha \rangle - \langle \mathbf{u}, v_\beta \rangle \geq \sqrt{1 - u_i^2} \cdot \mathrm{wid}_{\mathbf{u}'}(\Delta_P) - 2u_i|y_{i,i}|.$$

On the other hand, since the i-th coordinates of all vertices of P have absolute values at most $|y_{i,i}|$, we have

$$\mathrm{wid}_{\mathbf{u}}(P) \leq \sqrt{1 - u_i^2} \cdot \mathrm{wid}_{\mathbf{u}'}(P) + 2u_i|y_{i,i}|.$$

Furthermore, we have $u_i|y_{i,i}| < c_i \cdot \mathrm{wid}_{\mathbf{u}}(P) = (c_{i+1}/5) \cdot \mathrm{wid}_{\mathbf{u}}(P)$ by assumption and $\mathrm{wid}_{\mathbf{u}'}(\Delta_P) \geq c_{i+1} \cdot \mathrm{wid}_{\mathbf{u}'}(P)$ by the induction hypothesis. Using these four

inequalities, we deduce that

$$\begin{aligned}
\langle \mathbf{u}, v_\alpha \rangle - \langle \mathbf{u}, v_\beta \rangle &\geq \sqrt{1-u_i^2} \cdot \mathrm{wid}_{\mathbf{u}'}(\Delta_P) - 2u_i|y_{i,i}| \\
&\geq c_{i+1}\sqrt{1-u_i^2} \cdot \mathrm{wid}_{\mathbf{u}'}(P) - 2u_i|y_{i,i}| \\
&\geq c_{i+1}\mathrm{wid}_{\mathbf{u}}(P) - 2c_{i+1}u_i|y_{i,i}| - 2u_i|y_{i,i}| \\
&\geq c_{i+1}\mathrm{wid}_{\mathbf{u}}(P) - 4u_i|y_{i,i}| \\
&> c_{i+1}\mathrm{wid}_{\mathbf{u}}(P) - (4c_{i+1}/5) \cdot \mathrm{wid}_{\mathbf{u}}(P) \\
&= (c_{i+1}/5) \cdot \mathrm{wid}_{\mathbf{u}}(P) \\
&= c_i \cdot \mathrm{wid}_{\mathbf{u}}(P).
\end{aligned}$$

Since $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq \langle \mathbf{u}, v_\alpha \rangle - \langle \mathbf{u}, v_\beta \rangle$, we have $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_i \cdot \mathrm{wid}(P)$. In both of the cases, we have $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_i \cdot \mathrm{wid}(P)$. Therefore, we can use this induction argument to finally obtain the constant c_1 (note that c_1 is truly a constant as d is fixed), which satisfies $\mathrm{wid}_{\mathbf{u}}(\Delta_P) \geq c_1 \cdot \mathrm{wid}_{\mathbf{u}}(P)$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$. As a result, $\mathrm{wid}_{\mathbf{u}}(\Delta_P) = \Theta(\mathrm{wid}_{\mathbf{u}}(P))$ for all unit vectors $\mathbf{u} \in \mathbb{R}^d$. In particular, $\mathrm{wid}(\Delta_P) = \Theta(\mathrm{wid}(P))$. □

The idea used for constructing the witness simplex is standard, and was previously used to construct approximate minimum-volume bounding boxes [33]. In fact, the approximate minimum-volume bounding box constructed in [33] can also be used here as witness objects for approximating the expected width, because it approximates the directional width with respect to any direction. However, the resulting algorithm will have a higher time complexity, because there can be $\Omega(n^{2d})$ possible approximate minimum-volume bounding boxes of a SCH of n stochastic points (a bounding box is determined by $2d$ points), while the number of the possible witness simplices is $O(n^{d+1})$.

3.3.2 An $O(1)$-approximation algorithm

With the notion of witness simplex defined above, we now use the witness approach to establish an approximation algorithm for computing $\mathrm{wid}_{\mathcal{S}}$. The basic idea is similar to what we use for approximating $\mathrm{diam}_{\mathcal{S}}$. We define

$$\mathrm{wid}^*_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot \mathrm{wid}(\Delta_{\mathcal{CH}(R)}),$$

Lemma 27 implies $\mathrm{wid}^*_{\mathcal{S}} = \Theta(\mathrm{wid}_{\mathcal{S}})$. Thus, in order to approximate $\mathrm{wid}_{\mathcal{S}}$ within a constant factor, it suffices to compute $\mathrm{wid}^*_{\mathcal{S}}$. To compute $\mathrm{wid}^*_{\mathcal{S}}$ by directly using the above formula takes exponential time, as S has 2^n subsets. However, since $\Delta_{\mathcal{CH}(R)}$ must be a d-simplex with vertices in S, $\mathrm{wid}^*_{\mathcal{S}}$ can also be written as

$$\mathrm{wid}^*_{\mathcal{S}} = \sum_{\Delta \in \Gamma^d_S} \Pr(\Delta) \cdot \mathrm{wid}(\Delta), \tag{3.7}$$

where Γ^d_S is the set of all d-simplices in $\mathbb{R}^d$ whose vertices are (distinct) points in S and

$$\Pr(\Delta) = \Pr_{R \sim \mathcal{S}}[\Delta_{\mathcal{CH}(R)} = \Delta]$$

is the probability that the witness simplex of a SCH of $\mathcal{S}$ is Δ. Note that $|\Gamma^d_S| = O(n^{d+1})$, which is polynomial. So the above formula allows us to compute $\mathrm{wid}^*_{\mathcal{S}}$ in polynomial time, as long as we are able to compute $\Pr(\Delta)$ efficiently for each $\Delta \in \Gamma^d_S$. Fixing $\Delta \in \Gamma^d_S$, we now investigate how to compute $\Pr(\Delta)$. As argued before, we can recover the vertex list $(v_0, \ldots, v_d)$ of Δ. By the construction of Δ, $v_0, \ldots, v_d$ are points in S. For $i \in \{0, \ldots, d-1\}$, we denote by E_i the i-dim hyperplane in $\mathbb{R}^d$ through $v_0, \ldots, v_i$. We give the following criterion for checking if Δ is the witness simplex of a SCH of $\mathcal{S}$. For a hyperplane H (of any dimension) in $\mathbb{R}^d$ and two points $a, b \in \mathbb{R}^d$, we write $a \prec_H b$ if $\mathrm{dist}(a, H) < \mathrm{dist}(b, H)$, or $\mathrm{dist}(a, H) = \mathrm{dist}(b, H)$ and $a \prec b$.

Lemma 28. *For a realization R of $\mathcal{S}$, Δ is the witness simplex of $\mathcal{CH}(R)$ (i.e., $\Delta = \Delta_{\mathcal{CH}(R)}$) iff the following two conditions hold.*
(1) *R contains $v_0, \ldots, v_d$.*
(2) *R does not contain any point $a \in S$ satisfying $v_0 \prec a$ or $v_{i+1} \prec_{E_i} a$ for some $i \in \{0, \ldots, d-1\}$.*

Proof. Let R be a realization of $\mathcal{S}$, and set $C = \mathcal{CH}(R)$. The proof of the lemma is somewhat straightforward by using the definition of witness simplex. To see the "if" part, assume the two conditions in the lemma hold. Then v_0 must be the largest point in R under $\prec$-order, which must be a vertex of C. Furthermore, v_{i+1} must be a vertex of C (for it is the farthest from E_i and the points in S are in general position) which has the maximum distance to E_i (in addition, if there exists another vertex v of C having the same distance to E_i as v_{i+1}, then $v \prec v_{i+1}$). Thus, by

definition, $\Delta = \Delta_C$. To see the "only if" part, assume $\Delta = \Delta_C$. Then $v_0, \ldots, v_d$ are vertices of C and must be contained in R, which implies condition (1). Since $(v_0, \ldots, v_d)$ is the vertex list of Δ, v_0 is the largest vertex of C under $\prec$-order. Also, for any $i \in \{0, \ldots, d-1\}$, v_{i+1} is a vertex of C which has the maximum distance to E_i (in addition, if there exists another vertex v of C having the same distance to E_i as v_{i+1}, then $v \prec v_{i+1}$), so R cannot contain any point a with $v_{i+1} \prec_{E_i} a$. So we have condition (2). □

Using the above lemma, we can, in a straightforward way, compute $\Pr(\Delta)$ in linear time, just by multiplying the existence probabilities of $v_0, \ldots, v_d$ and the non-existence probabilities of all $a \in S$ which should not be included in R (according to condition (2) in the lemma). Therefore, we obtain an $O(n^{d+2})$-time algorithm for computing wid^*_S. It is easy to improve the runtime to $O(n^{d+1} \log n)$ as follows. We enumerate all $\Delta \in \Gamma^d_S$ by considering their vertex lists. Fixing d (distinct) points $v_0, \ldots, v_{d-1} \in S$, we show how to compute $\Pr(\Delta)$ for all $\Delta \in \Gamma^d_S$ whose vertex lists are of the form $(v_0, \ldots, v_{d-1}, \cdot)$ in $O(n \log n)$ time. First, we determine a subset $V \subseteq S \backslash \{v_0, \ldots, v_{d-1}\}$ consisting of all $v \in S \backslash \{v_0, \ldots, v_{d-1}\}$ such that $(v_0, \ldots, v_{d-1}, v)$ is the vertex list of the d-simplex whose vertices are $v_0, \ldots, v_{d-1}, v$. Clearly, this step can be completed in linear time by enumerating all $v \in S \backslash \{v_0, \ldots, v_{d-1}\}$ and verifying for each v whether $v \in V$. If $V = \emptyset$, we are done because there is no $\Delta \in \Gamma^d_S$ whose vertex list is of the form $(v_0, \ldots, v_{d-1}, \cdot)$. So suppose $V \neq \emptyset$. For $i \in \{0, \ldots, d-1\}$, we denote by E_i be the i-dim hyperplane in $\mathbb{R}^d$ through $v_0, \ldots, v_i$. We then compute a subset $A \subset S$ consisting of all $a \in S$ such that $v_0 \prec a$ or $v_{i+1} \prec_{E_i} a$ for some $i \in \{0, \ldots, d-2\}$. Now for any $v \in V$, we denote by B_v the set of all $b \in S \backslash A$ such that $v \prec_{E_{d-1}} b$. By Lemma 28, we have

$$\Pr(\Delta_v) = \left(\prod_{i=0}^{d-1} \pi(v_i) \cdot \prod_{a \in A} (1 - \pi(a)) \right) \cdot \left(\pi(v) \cdot \prod_{b \in B_v} (1 - \pi(b)) \right), \tag{3.8}$$

where Δ_v is the d-simplex with vertices $v_0, \ldots, v_{d-1}, v$. Note that the left side of the above formula is independent of v and thus only needs to be computed once. To compute the right side efficiently, suppose $S \backslash A = \{c_1, \ldots, c_r\}$. We relabel these points such that $c_1 \prec_{E_{d-1}} \cdots \prec_{E_{d-1}} c_r$. This can be done by sorting in $O(n \log n)$ time, or more precisely, $O(r \log r)$ time. We then compute $\prod_{j=i}^{r} (1 - \pi(c_j))$ for all

$i \in \{1, \ldots, r\}$ (note that this can be done in linear time). With this in hand, we consider each $v \in V$. Since $V \subseteq S \backslash A$, we must have $v = c_i$ for some $i \in \{1, \ldots, r\}$. In this case, the right side of Equation 3.8 is just $\pi(c_i) \cdot \prod_{j=i+1}^{r}(1 - \pi(c_j))$ and hence can be computed in constant time. Therefore, we can compute $\Pr(\Delta_v)$ for all $v \in V$ in linear time. Including the time for sorting, this gives us an $O(n^{d+1} \log n)$ time algorithm for computing $\mathrm{wid}^*_{\mathcal{S}}$, i.e., approximating $\mathrm{wid}_{\mathcal{S}}$ within a constant factor.

Theorem 29. *One can $O(1)$-approximate $\mathrm{wid}_{\mathcal{S}}$ in $O(n^{d+1} \log n)$ time. The constant approximation factor could be exponential in d.*

3.3.3 A fully polynomial-time randomized approximation scheme

In this section, we develop a fully polynomial-time randomized approximation scheme (FPRAS) for computing $\mathrm{wid}_{\mathcal{S}}$. An FPRAS should take $\mathcal{S}$ and a real number $\varepsilon > 0$ as input and should output a $(1 + \varepsilon)$-approximation of $\mathrm{wid}_{\mathcal{S}}$ in time polynomial in the size of $\mathcal{S}$ and $1/\varepsilon$ with probability at least $2/3$.

We first introduce some notations. As defined in the preceding section, Γ^d_S is the set of all d-simplices in $\mathbb{R}^d$ whose vertices are (distinct) points in S, and for each $\Delta \in \Gamma^d_S$ the notation $\Pr(\Delta)$ denotes the probability that the witness simplex of a SCH of $\mathcal{S}$ is Δ. Let R be a realization of $\mathcal{S}$ and $\Delta \in \Gamma^d_S$ be a simplex. From Lemma 28, we know that $\Delta = \Delta_{\mathcal{CH}(R)}$ iff R contains the vertices of Δ but does not contain some other points in S according to condition (2) in the lemma. We now use V_Δ to denote the set of the vertices of Δ, X_Δ to denote the set of the points in S that R must not contain if $\Delta = \Delta_{\mathcal{CH}(R)}$. Let $F_\Delta = S \backslash (V_\Delta \cup X_\Delta)$, which is the set of the points in S whose presence/absence in R does not influence whether $\Delta = \Delta_{\mathcal{CH}(R)}$. Define $\mathcal{F}_\Delta$ as the sub-dataset of $\mathcal{S}$ with the point-set F_Δ. Our FPRAS works as follows. First, for each $\Delta \in \Gamma^d_S$, we randomly generate $m = \gamma \log n / \varepsilon^2$ realizations of $\mathcal{F}_\Delta$, where γ is a large enough constant to be determined. Let $R^\Delta_1, \ldots, R^\Delta_m$ be the generated realizations of $\mathcal{F}_\Delta$, and set $T^\Delta_i = R^\Delta_i \cup V_\Delta$. Note that the witness simplex of $\mathcal{CH}(T^\Delta_i)$ is Δ by Lemma 28. We then compute

$$\mathrm{wid}'_{\mathcal{S}} = \sum_{\Delta \in \Gamma^d_S} \Pr(\Delta) \cdot \left(\sum_{i=1}^{m} \frac{\mathrm{wid}(\mathcal{CH}(T^\Delta_i))}{m} \right), \tag{3.9}$$

and output $\mathrm{wid}'_{\mathcal{S}}$ as the approximation of $\mathrm{wid}_{\mathcal{S}}$.

Next, we discuss the choice of the constant γ and verify the correctness of our FPRAS. By Lemma 27, we can find positive constants k_1, k_2 such that $k_1 \cdot \mathrm{wid}(\Delta_P) \leq \mathrm{wid}(P) \leq k_2 \cdot \mathrm{wid}(\Delta_P)$ for any convex polytope P in $\mathbb{R}^d$ with $\mathrm{wid}(P) > 0$. We set $\gamma = d(k_2/k_1)^2$. With this choice of γ, we claim the following, which shows the correctness of our FPRAS.

Lemma 30. $(1-\varepsilon)\mathrm{wid}_{\mathcal{S}} \leq \mathrm{wid}'_{\mathcal{S}} \leq (1+\varepsilon)\mathrm{wid}_{\mathcal{S}}$ *with probability at least* $2/3$.

Proof. Indeed, we can write

$$\mathrm{wid}_{\mathcal{S}} = \sum_{\Delta \in \Gamma_S^d} \Pr(\Delta) \cdot \mathbf{E}_\Delta,$$

where $\mathbf{E}_\Delta$ is the conditional expected width of a SCH of $\mathcal{S}$ under the condition that the witness simplex of the SCH is Δ. Since $\mathrm{wid}'_{\mathcal{S}}$ is computed using Equation 3.9, it suffices to show that

$$(1-\varepsilon)\mathbf{E}_\Delta \leq \sum_{i=1}^{m} \frac{\mathrm{wid}(\mathcal{CH}(T_i^\Delta))}{m} \leq (1+\varepsilon)\mathbf{E}_\Delta \tag{3.10}$$

for all $\Delta \in \Gamma_S^d$ with probability at least $2/3$. Fixing $\Delta \in \Gamma_S^d$, we can regard $\mathrm{wid}(\mathcal{CH}(T_1^\Delta)), \ldots, \mathrm{wid}(\mathcal{CH}(T_m^\Delta))$ as i.i.d. random variables. By Lemma 28 and the construction of each T_i^Δ, we know that the expectation of $\mathrm{wid}(\mathcal{CH}(T_i^\Delta))$ is $\mathbf{E}_\Delta$. Furthermore, we have $k_1 \cdot \mathrm{wid}(\Delta) \leq \mathrm{wid}(\mathcal{CH}(T_i^\Delta)) \leq k_2 \cdot \mathrm{wid}(\Delta)$, since the witness simplex of $\mathcal{CH}(T_i^\Delta)$ is Δ as argued before. Based on these observations, we can apply Hoeffding's inequality to obtain

$$\Pr\left[\left|\sum_{i=1}^{m} \frac{\mathrm{wid}(\mathcal{CH}(T_i^\Delta))}{m} - \mathbf{E}_\Delta\right| \geq \varepsilon \mathbf{E}_\Delta\right] \leq 2\exp\left(-\frac{2m \cdot (\varepsilon \mathbf{E}_\Delta)^2}{(k_2-k_1)^2 \cdot \mathrm{wid}(\Delta)^2}\right).$$

Note that $m = \gamma \log n/\varepsilon^2 = d(k_2/k_1)^2 \log n/\varepsilon^2$. Therefore,

$$-\frac{2m \cdot (\varepsilon \mathbf{E}_\Delta)^2}{(k_2-k_1)^2 \cdot \mathrm{wid}(\Delta)^2} \leq -2d \log n,$$

since $\mathbf{E}_\Delta \geq k_1 \cdot \mathrm{wid}(\Delta)$. It follows that Equation 3.10 fails with probability $O(n^{-2d})$ for a specific Δ. Therefore, by union bound, Equation 3.10 holds for all $\Delta \in \Gamma_S^d$ with probability $1 - O(n^{-d+1})$, which is greater than $2/3$ for large n (assume $d \geq 2$). As a result, the inequality in the theorem is proved. □

Theorem 31. *There exists an FPRAS for computing* $\mathrm{wid}_{\mathcal{S}}$.

3.3.4 A polynomial-time approximation scheme

In this section, we design a PTAS for computing $\text{wid}_{\mathcal{S}}$. The high-level strategy to design such a PTAS is similar to that in the expected-diameter problem (Section 3.2.3). We shall first establish a formula for $\text{wid}_{\mathcal{S}}$ using conditional expectation, and then compute the conditional expectations using the grid technique as in the last section. Consider a realization R of $\mathcal{S}$. For a simplex Δ in $\mathbb{R}^d$, define E_Δ as the event that the witness simplex of R is Δ. Let $\mathcal{E} = \{E_\Delta : \Delta \text{ is a simplex whose vertices are in } S\}$. Then we can write

$$\text{wid}_{\mathcal{S}} = \sum_{E_\Delta \in \mathcal{E}} \Pr_{R\sim\mathcal{S}}[E_\Delta] \cdot \mathbb{E}_{R\sim\mathcal{S}}[\text{wid}(R)|E_\Delta]. \tag{3.11}$$

Note that $|\mathcal{E}| = O(n^{d+1})$. As argued before, the probability $\Pr_{R\sim\mathcal{S}}[E_\Delta]$ can be computed in $O(n)$ time. Thus, it suffices to (approximately) compute $\mathbb{E}_{R\sim\mathcal{S}}[\text{wid}(R)|E_\Delta]$ for all $E_\Delta \in \mathcal{E}$. Fix $E_\Delta \in \mathcal{E}$. Similar to the approach used in the expected-diameter problem, we shall first build a stochastic dataset $\mathcal{S}_\Delta = (S_\Delta, \pi_\Delta)$ such that $\mathbb{E}_{R\sim\mathcal{S}}[\text{wid}(R)|E_\Delta] = \text{wid}_{\mathcal{S}_\Delta}$. Let $(v_0, \ldots, v_d)$ be the vertex list of Δ, and T_Δ be the subset of S consisting of all the points to the left of v_0 or farther from F_i than v_{i+1} for some $i \in \{0, \ldots, d-1\}$. We define $S_\Delta = S \backslash T_\Delta$ and

$$\pi_\Delta(a) = \begin{cases} 1 & \text{if } a \in \{v_0, \ldots, v_d\}, \\ \pi(a) & \text{otherwise.} \end{cases}$$

Clearly, E_Δ happens iff $v_0, \ldots, v_d$ exists and all the points in T_Δ do not exist. Thus, we have the equation $\mathbb{E}_{R\sim\mathcal{S}}[\text{wid}(R)|E_\Delta] = \text{wid}_{\mathcal{S}_\Delta}$. It suffices to have a PTAS for computing $\text{wid}_{\mathcal{S}_\Delta}$. To this end, we apply the grid technique used in the expected-diameter problem: building a grid and "compressing" the stochastic points in each cell. By [34], we can compute (in constant time) a bounding box B of Δ such that $\text{wid}_{\mathbf{u}}(B) = \Theta(\text{wid}_{\mathbf{u}}(\Delta))$ for all directions $\mathbf{u} \in \mathbb{S}^{d-1}$. Without loss of generality, we assume that $B = \prod_{i=1}^{d}[0, \delta_i]$. Now we build a grid Γ on $\mathbb{R}^d$ consisting of hyper-rectangle cells of size $(c \cdot \varepsilon\delta_1) \times \cdots \times (c \cdot \varepsilon\delta_d)$ where c is a sufficiently small constant and ε is the approximation factor. In other words, each cell $\square$ of Γ is an axis-parallel hyper-rectangle whose side-length in the i-th dimension is $c \cdot \varepsilon\delta_d$. For each cell $\square$, we denote by $c_\square$ as the center of $\square$. We construct a new stochastic dataset

$\mathcal{S}'_\Delta = (S'_\Delta, \pi'_\Delta)$ as $S'_\Delta = \{c_\square : S_\Delta \cap \square \neq \emptyset\}$ and

$$\pi'_\Delta(c_\square) = \prod_{a \in S_\Delta \cap \square} (1 - \pi_\Delta(a)).$$

In what follows, we shall bound the size of $\mathcal{S}'_\Delta$ and show that $\mathrm{wid}_{\mathcal{S}'_\Delta}$ is a good approximation of $\mathrm{wid}_{\mathcal{S}_\Delta}$. We first observe the following.

Lemma 32. *For all possible realizations R of $\mathcal{S}_\Delta$ and all directions $\mathbf{u} \in \mathbb{S}^{d-1}$, $\mathrm{wid}_{\mathbf{u}}(R) = \Theta(\mathrm{wid}_{\mathbf{u}}(B))$.*

Proof. As argued before, for a subset $S' \subseteq S$, the witness simplex of S' is Δ if $v_0, \ldots, v_d \in S'$ and $T_\Delta \cap S' = \emptyset$. Since $\pi_\Delta(a) = 1$ if $a \in \{v_0, \ldots, v_d\}$, any possible realization R of $\mathcal{S}_\Delta$ must contain $v_0, \ldots, v_d$. Furthermore, $T_\Delta \cap S_\Delta = \emptyset$. Therefore, the witness simplex of any possible realization R of $\mathcal{S}_\Delta$ is Δ. By Lemma 27, we have $\mathrm{wid}_{\mathbf{u}}(R) = \Theta(\mathrm{wid}_{\mathbf{u}}(\Delta))$ for all directions $\mathbf{u} \in \mathbb{S}^{d-1}$. Since the box B satisfies $\mathrm{wid}_{\mathbf{u}}(B) = \Theta(\mathrm{wid}_{\mathbf{u}}(\Delta))$ for all $\mathbf{u} \in \mathbb{S}^{d-1}$, the statement in the lemma holds. □

The above lemma implies $\mathrm{wid}_{\mathbf{u}}(S_\Delta) = \Theta(\mathrm{wid}_{\mathbf{u}}(B))$ for all $\mathbf{u} \in \mathbb{S}^{d-1}$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_d \in \mathbb{S}^{d-1}$ be the standard basis of $\mathbb{R}^d$. Then we have $\mathrm{wid}_{\mathbf{e}_i}(S_\Delta) = \Theta(\delta_i)$ for all $i \in \{1, \ldots, d\}$, which implies that the points in S_Δ are contained in an orthogonal box B' whose side-length in the i-th dimension is $\Theta(\delta_i)$. Since the side-length of each grid cell of Γ in the i-th dimension is $c \cdot \varepsilon\delta_i$, the number of the grid cells intersecting B is $O(\varepsilon^{-d})$. It follows that the number of the grid cells $\square$ satisfying $S_\Delta \cap \square \neq \emptyset$ is $O(\varepsilon^{-d})$, and hence $|S'_\Delta| = O(\varepsilon^{-d})$. To see that $\mathrm{wid}_{\mathcal{S}'_\Delta}$ is a good approximation of $\mathrm{wid}_{\mathcal{S}_\Delta}$, we establish a lemma similar to Lemma 22.

Lemma 33. $(1 - \varepsilon) \cdot \mathrm{wid}_{\mathcal{S}_\Delta} \leq \mathrm{wid}_{\mathcal{S}'_\Delta} \leq (1 + \varepsilon) \cdot \mathrm{wid}_{\mathcal{S}_\Delta}$.

Proof. Consider the map $f : S_\Delta \to S'_\Delta$ defined as $f(a) = c_\square$ where $\square$ is the cell containing a. A subset P of S_Δ is then mapped to a subset $P' = f(P)$ of S'_Δ. Let R be a possible realization of S_Δ and $R' = f(R)$. We first show that $1 - \varepsilon/2 \leq \mathrm{wid}(R')/\mathrm{wid}(R) \leq 1 + \varepsilon/2$. Let $\mathbf{u} \in \mathbb{S}^{d-1}$ be the direction defining $\mathrm{wid}(R')$, that is $\mathrm{wid}(R') = \mathrm{wid}_{\mathbf{u}}(R')$, and $a, a' \in R$ be the two points defining $\mathrm{wid}_{\mathbf{u}}(R)$, that is, $\mathrm{wid}_{\mathbf{u}}(R) = \mathrm{wid}_{\mathbf{u}}(\{a, a'\})$. Denote by $\square$ and $\square'$ the grid cells containing a and a', respectively. Then $f(a) = c_\square$ and $f(a') = c_{\square'}$. Note that

$$\mathrm{wid}_{\mathbf{u}}(\{a, a'\}) \leq \mathrm{wid}_{\mathbf{u}}(\{c_\square, c_{\square'}\}) + \mathrm{wid}_{\mathbf{u}}(\square) + \mathrm{wid}_{\mathbf{u}}(\square'),$$

and $\mathrm{wid}_{\mathbf{u}}(\square) = \mathrm{wid}_{\mathbf{u}}(\square') = c \cdot \varepsilon \mathrm{wid}_{\mathbf{u}}(B)$. It follows that

$$\frac{\mathrm{wid}(R')}{\mathrm{wid}(R)} \geq \frac{\mathrm{wid}_{\mathbf{u}}(R')}{\mathrm{wid}_{\mathbf{u}}(R)} \geq \frac{\mathrm{wid}_{\mathbf{u}}(\{c_{\square}, c_{\square'}\})}{\mathrm{wid}_{\mathbf{u}}(\{a, a'\})} \geq 1 - \frac{2c \cdot \varepsilon \mathrm{wid}_{\mathbf{u}}(B)}{\mathrm{wid}_{\mathbf{u}}(\{a, a'\})}.$$

By Lemma 32, we have $\mathrm{wid}_{\mathbf{u}}(R) = \Theta(\mathrm{wid}_{\mathbf{u}}(B))$. Since c is sufficiently small, we have $\mathrm{wid}(R')/\mathrm{wid}(R) \geq 1 - \varepsilon$. Using a similar argument, we can show that $\mathrm{wid}(R')/\mathrm{wid}(R) \leq 1 + \varepsilon$.

Based on this, we can complete the proof using the same approach as in the proof of Lemma 22. From the construction of $\mathcal{S}'_{\Delta}$, we have

$$\Pr_{R' \sim \mathcal{S}'_{\Delta}}[R' = P'] = \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}_{\Delta}}[R = P].$$

It follows that

$$\mathrm{wid}_{\mathcal{S}'_{\Delta}} = \sum_{P' \subseteq S'_{\Delta}} \Pr_{R' \sim \mathcal{S}'_{\Delta}}[R' = P'] \cdot \mathrm{wid}(P') = \sum_{P' \subseteq S'_{\Delta}} \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}_{\Delta}}[R = P] \cdot \mathrm{wid}(P').$$

On the other hand, we have

$$\mathrm{wid}_{\mathcal{S}_{\Delta}} = \sum_{P \subseteq S_{\Delta}} \Pr_{R \sim \mathcal{S}_{\Delta}}[R = P] \cdot \mathrm{wid}(P) = \sum_{P' \subseteq S'_{\Delta}} \sum_{P, f(P) = P'} \Pr_{R \sim \mathcal{S}_{\Delta}}[R = P] \cdot \mathrm{wid}(P).$$

As argued above, if $f(P) = P'$ and $\Pr_{R \sim \mathcal{S}_{\Delta}}[R = P] > 0$, then $1 - \varepsilon \leq \mathrm{wid}(P')/\mathrm{wid}(P) \leq 1 + \varepsilon$. Therefore, we have $(1 - \varepsilon) \cdot \mathrm{wid}_{\mathcal{S}_{\Delta}} \leq \mathrm{wid}_{\mathcal{S}'_{\Delta}} \leq (1 + \varepsilon) \cdot \mathrm{wid}_{\mathcal{S}_{\Delta}}$. □

The fact that $|S'_{\Delta}| = O(\varepsilon^{-d})$ allows us to compute $\mathrm{wid}_{\mathcal{S}'_{\Delta}}$ in $O(\varepsilon^{-d} \cdot 2^{\varepsilon^{-d}})$ time. By the above lemma, this results in a PTAS for computing $\mathrm{wid}_{\mathcal{S}_{\Delta}}$. Further combining this with Equation 3.11, we obtain a PTAS for computing $\mathrm{wid}_{\mathcal{S}}$.

Theorem 34. *There exists a PTAS for computing $\mathrm{wid}_{\mathcal{S}}$.*

3.4 Computing the expected combinatorial complexity

Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset in $\mathbb{R}^d$ with d fixed, and suppose $|S| = n$. Our goal in this section is to (exactly) compute the expected complexity of a SCH of $\mathcal{S}$, defined as

$$\mathrm{comp}_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot |\mathcal{CH}(R)|,$$

where $\Pr(R)$ denotes the probability that R occurs as a realization of $\mathcal{S}$.

3.4.1 Reduction to SCH membership probability queries

Given a stochastic dataset $\mathcal{T}$ in $\mathbb{R}^d$ and a query point $q \in \mathbb{R}^d$, the SCH membership probability (of q with respect to $\mathcal{T}$) refers to the probability that q lies in a SCH of $\mathcal{T}$, which we denote by $\text{mem}_{\mathcal{T}}(q)$. It is known that $\text{mem}_{\mathcal{T}}(q)$ can be computed in $O(m^{d-1})$ time for $d \geq 3$ [21, 22] and $O(m \log m)$ time for $d \in \{1, 2\}$ [7], where m is the number of the stochastic points in $\mathcal{T}$.

In this section, we reduce our problem of computing $\text{comp}_{\mathcal{S}}$ to SCH membership probability queries. Let R be a realization of $\mathcal{S}$. It is clear that the faces of $\mathcal{CH}(R)$ must be simplices with vertices in S. Therefore, we can rewrite the formula for $\text{comp}_{\mathcal{S}}$ as

$$\text{comp}_{\mathcal{S}} = \sum_{R \subseteq S} \Pr(R) \cdot \left(\sum_{\Delta \in \Gamma_S} \sigma(R, \Delta) \right) = \sum_{\Delta \in \Gamma_S} F_\Delta, \tag{3.12}$$

where Γ_S is the set of all simplices (of dimension less than d) with vertices in S, σ is an indicator function such that $\sigma(R, \Delta) = 1$ if Δ is a face of $\mathcal{CH}(R)$ and $\sigma(R, \Delta) = 0$ otherwise, and F_Δ is the probability that Δ is a face of a SCH of $\mathcal{S}$. We now show that for each $\Delta \in \Gamma_S$, the computation of F_Δ can be reduced to a SCH membership probability query. Suppose Y is a set of m ($m \geq d + 1$) points in $\mathbb{R}^d$ in general position. Let $y_0, \ldots, y_k \in Y$ be $k + 1$ points where $0 \leq k \leq d - 1$, and Δ be the k-simplex with vertices $y_0, \ldots, y_k$. Define vectors $\mathbf{u}_i = y_i - y_0$ for $i \in \{1, \ldots, k\}$. By the general position assumption, $\mathbf{u}_1, \ldots, \mathbf{u}_k$ generate a k-dim linear subspace H of $\mathbb{R}^d$. Set H^* to be the orthogonal complement of H in $\mathbb{R}^d$, which is by definition the $(d - k)$-dim linear subspace of $\mathbb{R}^d$ orthogonal to H. We then orthogonally project the points in Y to H^*, and denote the set of the projection images by Y^*. Note that $y_0, \ldots, y_k$ are clearly projected to the same point in H^*, which we denote by $\hat{y}$. We have the following observation.

Lemma 35. *Δ is a face of $\mathcal{CH}(Y)$ iff $\hat{y}$ is a vertex of $\mathcal{CH}(Y^*)$ in H^*.*

Proof. Suppose $Y = \{y_0, y_1, \ldots, y_m\}$, and let $P = \mathcal{CH}(Y)$, $P^* = \mathcal{CH}(Y^*)$. Then any point $x \in P$ can be represented as a linear combination $x = \sum_{i=0}^{m} w_i \cdot y_i$ where $w_i \geq 0$ and $\sum_{i=0}^{m} w_i = 1$, which we call *convex representation*. It is easy to check that x is on the boundary of P iff x has a unique convex representation and in which there are at most d nonzero w_i's. We first show the "if" part. Assume Δ is not a

face of $\mathcal{CH}(Y)$. Then there must exist $x \in \Delta$ which is not on the boundary of P. Since Δ is a simplex, there is a unique convex representation of x satisfying $w_i = 0$ for all $i > k$. But this should not be the only convex representation of x, because x is not on the boundary of P. Therefore, x has another convex representation with $w_i > 0$ for some $i > k$ (without loss of generality, assume $w_m > 0$). Let $\rho : \mathbb{R}^d \to H^*$ be the orthogonal projection map. We have

$$\hat{y} = \rho(x) = \rho\left(\sum_{i=0}^{m} w_i \cdot y_i\right) = \sum_{i=0}^{m} w_i \cdot \rho(y_i).$$

Note that all $\rho(y_i)$ are points in P^*. Furthermore, by general position assumption, $\rho(y_m) \neq \hat{y}$. Therefore, $\hat{y}$ is not a vertex of P^*. Next, we consider the "only if" part. Assume $\hat{y}$ is not a vertex of P^*. Then we have $P^* = \mathcal{CH}(Y^* \backslash \{\hat{y}\})$. It follows that $\hat{y}$ has a convex representation $\hat{y} = \sum_{i=0}^{m} w_i \cdot \rho(y_i)$ with $w_0 = \cdots = w_k = 0$. Lifting this representation, we obtain a point $x = \sum_{i=0}^{m} w_i \cdot y_i \in P$. Since $\rho(x) = \hat{y}$, x is in the k-dim hyperplane L spanned by $y_0, \ldots, y_k$. Now assume Δ is a face of P, so we must have $L \cap P = \Delta$, which implies $x \in \Delta$. This means that x has a convex representation with $w_{k+1} = \cdots = w_m = 0$. Since x has two different convex representations, it is not on the boundary of P, contradicting that $x \in \Delta$. As a result, Δ is not a face of P. □

By the above lemma, we can reduce the computation of F_Δ for any $\Delta \in \mathit{\Gamma}_S$ to a SCH membership query as follows. For each $i \in \{0, \ldots, d-1\}$, let $\mathit{\Gamma}_S^i \subseteq \mathit{\Gamma}_S$ be the subset consisting of all i-simplices in $\mathit{\Gamma}_S$ (then $\mathit{\Gamma}_S = \bigcup_{i=0}^{d-1} \mathit{\Gamma}_S^i$). Suppose $\Delta \in \mathit{\Gamma}_S^k$ is a k-simplex with vertices $v_0, \ldots, v_k \in S$. As before, we define vectors $\mathbf{u}_i = v_i - v_0$ for $i \in \{1, \ldots, k\}$. Then $\mathbf{u}_1, \ldots, \mathbf{u}_k$ generate a k-dim linear subspace H of $\mathbb{R}^d$, and set H^* to be the orthogonal complement of H in $\mathbb{R}^d$. Let $\rho : \mathbb{R}^d \to H^*$ be the orthogonal projection map. We define a multi-set $S' = \{\rho(a) : a \in S \backslash \{v_0, \ldots, v_k\}\}$ of points in H^*, which in turn gives us a stochastic dataset $\mathcal{S}' = (S', \pi')$ in H^* where $\pi'(\rho(a)) = \pi(a)$. Set $q = \rho(v_0) = \cdots = \rho(v_k)$.

Corollary 36. $F_\Delta = \prod_{i=0}^{k} \pi(v_i) \cdot (1 - \mathrm{mem}_{\mathcal{S}'}(q))$.

Proof. Let R be a realization of $\mathcal{S}$. If Δ is a face of $\mathcal{CH}(R)$, then $v_0, \ldots, v_k$ must

be contained in R. Furthermore, by Lemma 35, q must be a vertex of the projection image of $\mathcal{CH}(R)$ in H^*. By the general position assumption, this is equivalent to saying that q is outside the projection image of $\mathcal{CH}(R\backslash\{v_0,\ldots,v_k\})$. Conversely, if $v_0,\ldots,v_k$ are contained in R and q is outside the projection image of $\mathcal{CH}(R\backslash\{v_0,\ldots,v_k\})$, then Δ is a face of $\mathcal{CH}(R)$ by Lemma 35. The probability that R contains $v_0,\ldots,v_k$ is $\prod_{i=0}^{k}\pi(v_i)$, and the probability that q is outside the projection image of $\mathcal{CH}(R\backslash\{v_0,\ldots,v_k\})$ is $1-\mathrm{mem}_{\mathcal{S}'}(q)$. These two events are clearly independent. Therefore, we have the formula in the corollary. □

Since H^* is linearly homeomorphic to $\mathbb{R}^{d-k}$, computing $\mathrm{mem}_{\mathcal{S}'}(q)$ is nothing but answering a SCH membership probability query in $\mathbb{R}^{d-k}$. Therefore, using the algorithms for answering SCH membership probability queries [21, 22], F_Δ can be computed in $O(n^{d-k-1})$ time if $k\in\{0,\ldots,d-3\}$. Note that $|\Gamma_S^k|=O(n^{k+1})$, so we can compute the sum $\sum_{i=0}^{d-3}\sum_{\Delta\in\Gamma_S^i}F_\Delta$ in $O(n^d)$ time. In order to further compute $\mathrm{comp}_{\mathcal{S}}$ by Equation 3.12, we now only need to compute $\sum_{\Delta\in\Gamma_S^{d-2}}F_\Delta$ and $\sum_{\Delta\in\Gamma_S^{d-1}}F_\Delta$. Answering SCH membership probability queries in $\mathbb{R}^1$ and $\mathbb{R}^2$ requires $O(m\log m)$ time [7] (where m is the size of the given stochastic dataset). Thus, if we use the algorithm in [7] to calculate SCH membership probabilities, our computation task cannot be done in $O(n^d)$ time. The next section discusses how to handle this issue.

3.4.2 Handling the cases $k=d-2$ and $k=d-1$

Set $\lambda_1=\sum_{\Delta\in\Gamma_S^{d-1}}F_\Delta$ and $\lambda_2=\sum_{\Delta\in\Gamma_S^{d-2}}F_\Delta$. For simplicity of exposition, we first fix a point $o\in\mathbb{R}^d$ such that $S\cup\{o\}$ is in general position. For every hyperplane E with $o\notin E$, we denote by E^+ the connected component of $\mathbb{R}^d\backslash E$ containing o, and by E^- the other one. Define the $\mathcal{S}$-*statistic* of E as a 3-tuple $\mathrm{stat}_{\mathcal{S}}(E)=(p^+,p^-,A)$ where $p^+=\prod_{a\in S\cap E^+}(1-\pi(a))$, $p^-=\prod_{a\in S\cap E^-}(1-\pi(a))$, $A=S\cap E$. Let $\mathcal{E}$ be the collection of the hyperplanes in $\mathbb{R}^d$ which go through exactly d points in S. Since $S\cup\{o\}$ is in general position, $\mathrm{stat}(E)$ is defined for every $E\in\mathcal{E}$. We say an algorithm computes the $\mathcal{S}$-statistics for $\mathcal{E}$ if it reports $\mathrm{stat}_{\mathcal{S}}(E)$ for all $E\in\mathcal{E}$ in an arbitrary order (without repetition).

Lemma 37. *If there exists an algorithm computing the $\mathcal{S}$-statistics for $\mathcal{E}$ in $O(t(n))$ time and $O(s(n))$ space, then one can compute λ_1 and λ_2 in $O(t(n))$ time and*

$O(s(n))$ space.

Proof. We first consider the computation of λ_1. Let $\Delta \in \Gamma_S^{d-1}$ and $E \in \mathcal{E}$ be the hyperplane through the d vertices of Δ. Suppose q and $\mathcal{S}'$ are the point and the stochastic dataset defined in Corollary 36 for computing F_Δ. Since $\mathrm{mem}_{\mathcal{S}'}(q)$ is a SCH membership query in $\mathbb{R}^1$, it is clear that $1 - \mathrm{mem}_{\mathcal{S}'}(q) = p^+ + p^- - p^+p^-$ if $\mathrm{stat}(E) = (p^+, p^-, A)$. Hence F_Δ can be computed from $\mathrm{stat}_{\mathcal{S}}(E)$ in constant time. Consider the algorithm provided for computing the $\mathcal{S}$-statistics for $\mathcal{E}$. At every time it reports some $\mathrm{stat}_{\mathcal{S}}(E) = (p^+, p^-, A)$, we use it to compute the corresponding F_Δ (note that Δ can be recovered from A) in constant time. By summing up all F_Δ, we obtain λ_1, which is done in $O(t(n))$ time and $O(s(n))$ space. To consider λ_2, we need a careful analysis of the witness-edge method in [7] for computing SCH membership probability in $\mathbb{R}^2$. Let $\mathcal{T} = (T, \tau)$ be a stochastic dataset in $\mathbb{R}^2$, and $q \in \mathbb{R}^2$ be a query point. The witness-edge method computes $1 - \mathrm{mem}_{\mathcal{T}}(q)$ as a summation in which the summands correspond one-to-one to the hyperplanes (i.e., lines) that go through q and one point in T. Furthermore, the summand corresponding to a hyperplane E can be computed from $\mathrm{stat}_{\mathcal{T}}(E)$ in constant time. See [7] for the details. Now we consider the computation of λ_2. Let $\Delta \in \Gamma_S^{d-2}$. Suppose q and $\mathcal{S}'$ are the point and the stochastic dataset defined in Corollary 36 for computing F_Δ. We can regard $(\mathcal{S}', q)$ as a SCH membership probability query in $\mathbb{R}^2$. Thus, by our observation about the witness-edge method and Corollary 36, F_Δ can be expressed as a summation with summands one-to-one corresponding to the lines through q and one point in the point-set of $\mathcal{S}'$ (we denote by $\mathcal{L}$ the collection of these lines). Note that there is also an one-to-one correspondence between $\mathcal{L}$ and a sub-collection $\mathcal{E}_\Delta \subset \mathcal{E}$ containing the hyperplanes through all the $d-1$ vertices of Δ. Moreover, $\mathrm{stat}_{\mathcal{S}'}(L)$ for $L \in \mathcal{L}$ can be recovered from $\mathrm{stat}_{\mathcal{S}}(E)$ for $E \in \mathcal{E}_\Delta$ corresponding to L in constant time. Therefore, we may charge each summand of F_Δ to the corresponding hyperplane $E \in \mathcal{E}_\Delta$. Now consider the algorithm provided for computing the $\mathcal{S}$-statistics for $\mathcal{E}$. At every time it reports $\mathrm{stat}_{\mathcal{S}}(E)$ for some $E \in \mathcal{E}$, we use it to compute all summands charged to E. Note that each $E \in \mathcal{E}$ belongs to exactly $d-1$ $\mathcal{E}_\Delta$'s, and hence is charged with exactly $d-1$ summands. Therefore, this computation can be done in constant time. By summing up all summands charged to all $E \in \mathcal{E}$, we finally obtain λ_2, which is done in $O(t(n))$ time

and $O(s(n))$ space. □

By the above lemma, it now suffices to establish an efficient algorithm for computing the $\mathcal{S}$-statistics for $\mathcal{E}$. We do this in the next section.

3.4.3 Computing the $\mathcal{S}$-statistics for $\mathcal{E}$

We describe an algorithm which computes the $\mathcal{S}$-statistics for $\mathcal{E}$ in $O(n^d)$ time and $O(n)$ space. Suppose $S = \{a_1, \ldots, a_n\}$. Then every hyperplane $E \in \mathcal{E}$ can be uniquely represented as a d-tuple $(a_{i_1}, \ldots, a_{i_d})$ where $a_{i_1}, \ldots, a_{i_d}$ are the points on E and $i_1 < \cdots < i_d$. We first describe an algorithm using $O(n^d \log n)$ time and $O(n)$ space. Fixing $d-1$ points $a_{i_1}, \ldots, a_{i_{d-1}} \in S$ with $i_1 < \cdots < i_{d-1}$, we show how to report, in $O(n \log n)$ time and $O(n)$ space, the $\mathcal{S}$-statistics of all hyperplanes (i.e., lines) in $\mathcal{E}$ which are represented as the form $(a_{i_1}, \ldots, a_{i_{d-1}}, \cdot)$. Define Y as the $(d-2)$-dim hyperplane in $\mathbb{R}^d$ spanned by $a_{i_1}, \ldots, a_{i_{d-1}}$. Let Z be the (unique) *vertical* $(d-1)$-dim hyperplane containing Y (by "vertical" we mean that Z is perpendicular to the hyperplane $x_d = 0$), and $\mathcal{E}' \subseteq \mathcal{E}$ be the sub-collection consisting of all hyperplanes in $\mathcal{E}$ which contain Y. Note that $|\mathcal{E}'| = n - d + 1$. We then sort the hyperplanes in $\mathcal{E}'$ in the *rotation order* around Y, that is, we assign to each hyperplane $E \in \mathcal{E}'$ a key value equal to the rotation angle from Z to E (the rotation is taken around Y with a fixed direction), and sort the lines by their key values. Assume $E_1, \ldots, E_{n-d+1}$ is the sorted list. Observe that $\text{stat}(E_{j+1})$ can be computed in constant time if $\text{stat}(E_j)$ is already in hand, basically because the points on each side of E_{j+1} are almost the same as those on each side of E_j except two points. By this observation, we may compute the $\mathcal{S}$-statistics of $E_1, \ldots, E_{n-d+1}$ in $O(n)$ time. Once $\text{stat}(E_j)$ is computed, we report it if E_j is represented as the form $(a_{i_1}, \ldots, a_{i_{d-1}}, \cdot)$. In this way, we obtain an $O(n^d \log n)$-time and $O(n)$-space algorithm.

To eliminate the $\log n$ factor in the time bound, we need to further apply the techniques of duality and topological sweep [32]. This approach heavily relies on an idea in [21] (which was used to improve the algorithm for computing the separability-probability), so here we only provide a sketch. Instead of fixing $d-1$ points, we fix $d-2$ points $a_{i_1}, \ldots, a_{i_{d-2}} \in S$ with $i_1 < \cdots < i_{d-2}$, and want to report, in $O(n^2)$ time and $O(n)$ space, $\text{stat}(E)$ for all $E \in \mathcal{E}$ which are represented as the

form $(a_{i_1}, \ldots, a_{i_{d-2}}, \cdot, \cdot)$. Note that if this can be done, we immediately obtain an $O(n^d)$-time and $O(n)$-space algorithm. Consider the point-set S in the dual space of $\mathbb{R}^d$. Every point $a_i \in S$ is dual to a $(d-1)$-dim hyperplane a_i^* in the dual space. Furthermore, a $(k-1)$-dim hyperplane spanned by k (distinct) points $a_{j_1}, \ldots, a_{j_k} \in S$ is dual to a $(d-k)$-dim hyperplane in the dual space, which is in fact the intersection of $a_{j_1}^*, \ldots, a_{j_k}^*$. Let D be the $(d-3)$-dim hyperplane spanned by $a_{i_1}, \ldots, a_{i_{d-2}}$, which is dual to a 2-dim hyperplane (i.e., a plane) D^* in the dual space. For each $a_i \in S \backslash \{a_{i_1}, \ldots, a_{i_{d-2}}\}$, the intersection of a_i^* and D^* is a line in D^* (which should be the dual of the $(d-2)$-dim hyperplane spanned by $a_{i_1}, \ldots, a_{i_{d-2}}, a_i$). These $n-d+2$ lines form a line arrangement in D^*. Suppose l_i^* is the line corresponding to a_i. In the line arrangement, there are $n-d+1$ intersection points on l_i^*, each of which is the dual of a hyperplane through $a_{i_1}, \ldots, a_{i_{d-2}}, a_i$ in the original space. The order of these intersection points appearing on l_i^* is just the rotation order of the corresponding hyperplanes in the original space. Therefore, if these intersection points are already sorted, we can compute the $\mathcal{S}$-statistic of each of the corresponding hyperplanes in amortized $O(1)$ time. But we cannot use sorting, as it takes $O(n \log n)$ time per line and we have $O(n)$ lines in the arrangement. Instead, we use topological sweep to visit the intersection points in the arrangement. In the process of topological sweep, the intersection points on each line is visited in order along the line (though not consecutively). When the first intersection point on a line is visited, we use brute-force to compute the $\mathcal{S}$-statistic of the corresponding hyperplane in $O(n)$ time. Then when we go to the next intersection point on the line, we can compute the $\mathcal{S}$-statistic of the corresponding hyperplane in constant time from the $\mathcal{S}$-statistic of the hyperplane corresponding to the previous intersection point. Once a $\mathcal{S}$-statistic is computed, we report it if the hyperplane is represented as the form $(a_{i_1}, \ldots, a_{i_{d-2}}, \cdot, \cdot)$. The topological sweep takes $O(n^2)$ time and $O(n)$ space. Thus, we obtain an algorithm computing the $\mathcal{S}$-statistics for $\mathcal{E}$ in $O(n^d)$ time and $O(n)$ space.

With the above algorithm in hand, Lemma 37 implies that we can compute λ_1 and λ_2 in $O(n^d)$ time and $O(n)$ space. By further combining this with what we have in Section 3.4.1, we can finally conclude the following.

Theorem 38. *One can compute the exact value of* $\mathrm{comp}_{\mathcal{S}}$ *in* $O(n^d)$ *time.*

Chapter 4

Stochastic dominance problems

Let $\mathcal{S} = (S, \mathrm{cl}, \pi)$ be a given colored stochastic dataset in $\mathbb{R}^d$ where $S = \{a_1, \ldots, a_n\}$. In this chapter, we study the CSD problem and the FBCSD problem for $\mathcal{S}$; see Section 1.1 for the statement of these problems. For convenience, throughout this section, when denoting a colored dataset $\mathcal{T} = (T, \mathrm{cl})$, we simply use the notation T if the color function cl is clear.

4.1 Preliminaries

We formally define some notions about the dominance relation. We say a point $p \in \mathbb{R}^d$ *dominates* another point $q \in \mathbb{R}^d$ (denoted by $p \succ q$) if the coordinate of p is greater than or equal to the coordinate of q in every dimension. In a colored dataset $\mathcal{T} = (T, \mathrm{cl})$, an *inter-color dominance* is a pair (a, b) of points in T such that $\mathrm{cl}(a) \neq \mathrm{cl}(b)$ and $a \succ b$. By naturally generalizing the conventional dominance relation, one can define the dominance relation with respect to a specific orthogonal basis of $\mathbb{R}^d$. Specifically, a point $p \in \mathbb{R}^d$ *dominates* another point $q \in \mathbb{R}^d$ with respect to an orthogonal basis $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ of $\mathbb{R}^d$ (denoted by $p \succ_B q$) if $\langle \mathbf{b}_i, p \rangle \geq \langle \mathbf{b}_i, q \rangle$ for all $i \in \{1, \ldots, d\}$, where $\langle \cdot, \cdot \rangle$ is the inner product. With this generalized definition, the conventional dominance relation is just the dominance relation with respect to the standard basis $E = (\mathbf{e}_1, \ldots, \mathbf{e}_d)$ of $\mathbb{R}^d$.

4.2 The colored stochastic dominance problem

Define $\Lambda_{\mathcal{S}}$ as the probability that a realization of $\mathcal{S}$ contains inter-color dominances. Set $\Gamma_{\mathcal{S}} = 1 - \Lambda_{\mathcal{S}}$, which is the *inter-color dominance-free probability*, i.e., the probability that a realization of $\mathcal{S}$ contains no inter-color dominances. The goal of the CSD problem is to compute $\Lambda_{\mathcal{S}}$ (or $\Gamma_{\mathcal{S}}$).

4.2.1 An algorithm for $d = 2$

The naïve method for solving the CSD problem is to enumerate all subsets of S and "count" those containing inter-color dominances. However, it requires exponential time, as there are $2^{|S|}$ subsets of S to be considered. In this section, we show that the CSD problem in $\mathbb{R}^2$ can be solved much more efficiently. Specifically, we propose an $O(n^2 \log^2 n)$-time algorithm to compute $\Gamma_{\mathcal{S}}$. For simplicity, we assume that the points in S have distinct x-coordinates and y-coordinates (if this is not the case, we can first "regularize" S as shown later in Lemma 55).

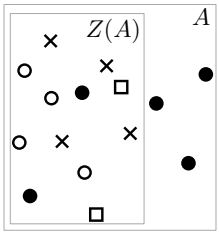


Figure 4.1: Illustrating A and $Z(A)$.

When computing $\Gamma_{\mathcal{S}}$, we need to consider the realizations which contain no inter-color dominances. As we will see, in the case where $d = 2$, these realizations have good properties, which allows us to solve the problem efficiently in a recursive way. For any point $a \in \mathbb{R}^2$, we use $x(a)$ (resp., $y(a)$) to denote the x-coordinate (resp., y-coordinate) of a. Suppose the points $a_1, \ldots, a_n \in S$ are already sorted such that $x(a_1) < \cdots < x(a_n)$. For convenience of exposition, we add a dummy point a_0 to S with $x(a_0) < x(a_1)$ and $y(a_0) > y(a_i)$ for all $i \in \{1, \ldots, n\}$. The color $\mathrm{cl}(a_0)$ is defined to be different from $\mathrm{cl}(a_1), \ldots, \mathrm{cl}(a_n)$, and $\pi(a_0) = 1$. Note that including a_0 does not change $\Gamma_{\mathcal{S}}$. For a subset $A = \{a_{i_1}, \ldots, a_{i_r}\}$ of S with $i_1 < \cdots < i_r$, we define $Z(A) = \emptyset$ if A is monochromatic, and otherwise $Z(A) = \{a_{i_1}, \ldots, a_{i_l}\}$

such that $\mathrm{cl}(a_{i_l}) \neq \mathrm{cl}(a_{i_{l+1}}) = \cdots = \mathrm{cl}(a_{i_r})$. In other words, $Z(A)$ is the subset of A obtained by dropping the "rightmost" points of the same color as a_{i_r}; see Figure 4.1. We have the following important observation.

Lemma 39. *A realization R of $\mathcal{S}$ contains no inter-color dominances iff $Z(R)$ contains no inter-color dominances and for any $a \in Z(R)$, $b \in R \backslash Z(R)$ it holds that $y(a) > y(b)$.*

Proof. To see the "if" part, assume that $Z(R)$ contains no inter-color dominances and $y(a) > y(b)$ for any $a \in Z(R)$, $b \in R \backslash Z(R)$. In this case, any two points in $Z(R)$ cannot form an inter-color dominance. Also, any two points in $R \backslash Z(R)$ cannot form an inter-color dominance for $R \backslash Z(R)$ is monochromatic. It suffices to show that any $a \in Z(R)$ and $b \in R \backslash Z(R)$ cannot form an inter-color dominance. By assumption, we have $y(a) > y(b)$. But by the definition of $Z(S)$, we also have $x(a) < x(b)$. Thus, a and b do not dominate each other. To see the "only if" part, assume R contains no inter-color dominances. Since $Z(R)$ is a subset of R, it also contains no inter-color dominances. Let $a \in Z(R)$ and $b \in R \backslash Z(R)$ be two points. As argued before, we have $x(a) < x(b)$. If $\mathrm{cl}(a) \neq \mathrm{cl}(b)$, then it is clear that $y(a) > y(b)$ (otherwise (a, b) forms an inter-color dominance). The only remaining case is $\mathrm{cl}(a) = \mathrm{cl}(b)$. Since $a \in Z(R)$, by the definition of $Z(R)$, we may find a point $o \in Z(R)$ such that $x(a) < x(o) < x(b)$ and $\mathrm{cl}(o) \neq \mathrm{cl}(a) = \mathrm{cl}(b)$. If $y(a) < y(b)$, then either $y(a) < y(o)$ or $y(o) < y(b)$, i.e., either (a, o) or (o, b) forms an inter-color dominance. Because R contains no inter-color dominances, we must have $y(a) > y(b)$. □

With this in hand, we then consider how to compute $\Gamma_{\mathcal{S}}$. For a nonempty subset $A \subseteq S$, we define the *signature*, $\mathrm{sgn}(A)$, of A as a pair (i, j) such that $a_i, a_j \in A$ and a_i (resp., a_j) has the greatest x-coordinate (resp., smallest y-coordinate) among all points in A. Let $E_{i,j}$ be the event that a realization R of $\mathcal{S}$ contains no inter-color dominances and satisfies $\mathrm{sgn}(R) = (i, j)$. Note that if a realization R contains no inter-color dominances, then either $R = \{a_0\}$ or some $E_{i,j}$ happens for $i, j \in \{1, \ldots, n\}$. So we immediately have

$$\Gamma_{\mathcal{S}} = \prod_{i=1}^{n} (1 - \pi(a_i)) + \sum_{i=1}^{n} \sum_{j=1}^{n} \Pr[E_{i,j}].$$

Now the problem is reduced to computing all $\Pr[E_{i,j}]$. Instead of working on the events $\{E_{i,j}\}$ directly, we consider a set of slightly different events $\{E'_{i,j}\}$ defined as follows. For $p \in \{0, \ldots, n\}$, set $S_p = \{a_0, \ldots, a_p\}$, and we use $\mathcal{S}_p$ to denote the sub-dataset of $\mathcal{S}$ with point set $S_p \subseteq S$. Define $E'_{i,j}$ as the event that a realization R of $\mathcal{S}_i$ contains no inter-color dominances and satisfies $\mathrm{sgn}(R) = (i, j)$. It is quite easy to see the equations

$$\Pr[E_{i,j}] = \Pr[E'_{i,j}] \cdot \prod_{t=i+1}^{n} (1 - \pi(a_t)).$$

Set $F(i, j) = \Pr[E'_{i,j}]$. We show how to compute all $F(i, j)$ recursively by applying Lemma 39. Observe that $F(i, j) = 0$ if $x(a_i) < x(a_j)$ (equivalently, $i < j$) or $y(a_i) < y(a_j)$ or $\mathrm{cl}(a_i) \neq \mathrm{cl}(a_j)$. Thus, it suffices to compute all $F(i, j)$ with $i \geq j$, $y(a_i) \geq y(a_j)$, $\mathrm{cl}(a_i) = \mathrm{cl}(a_i)$ (we say the pair (i, j) is *legal* if these three conditions hold). Let (i, j) be a legal pair. Trivially, for $i = j = 0$, we have $F(i, j) = 1$. So suppose $i, j > 0$. Let R be a realization of $\mathcal{S}_i$. To compute $F(i, j)$, we consider the signature $\mathrm{sgn}(Z(R))$ under the condition that $E'_{i,j}$ happens. First, when $E'_{i,j}$ happens, we always have $Z(R) \neq \emptyset$, because R at least contains a_0, a_i, a_j (possibly $i = j$) and $\mathrm{cl}(a_0) \neq \mathrm{cl}(a_i) = \mathrm{cl}(a_j)$. Therefore, in this case, $\mathrm{sgn}(Z(R))$ is defined and must be a legal pair (i', j') for some $i', j' \in \{0, \ldots, i-1\}$. It follows that $F(i, j)$ can be computed by considering for each such pair (i', j') the probability that R contains no inter-color dominances and $\mathrm{sgn}(R) = (i, j)$, $\mathrm{sgn}(Z(R)) = (i', j')$, and then summing up these probabilities. Note that if $\mathrm{sgn}(R) = (i, j)$ and $\mathrm{sgn}(Z(R)) = (i', j')$, then $i' < j$ and $\mathrm{cl}(i') \neq \mathrm{cl}(i)$. In addition, if R contains no inter-color dominances, then we must have $y(a_i) < y(a_{j'})$ by Lemma 39. As such, we only need to consider the legal pairs (i', j') satisfying $i' < j$, $y(a_i) < y(a_{j'})$, $\mathrm{cl}(i') \neq \mathrm{cl}(i)$ (we denote the set of these pairs by $J_{i,j}$). Fixing such a pair $(i', j') \in J_{i,j}$, we investigate the corresponding probability. By the definition of $Z(R)$ and Lemma 39, we observe that if R contains no inter-color dominances and $\mathrm{sgn}(R) = (i, j)$, $\mathrm{sgn}(Z(R)) = (i', j')$, then

- $R \cap S_{i'}$ contains no inter-color dominances and $\mathrm{sgn}(R \cap S_{i'}) = (i', j')$;
- $R \cap (S_i \backslash S_{i'})$ includes a_i and a_j, but does not include any point a_t for $t \in \{i' + 1, \ldots, i\}$ satisfying $\mathrm{cl}(a_t) \neq \mathrm{cl}(a_i)$ or $y(a_t) < y(a_j)$ or $y(a_{j'}) < y(a_t)$.

Conversely, one can also verify that if a realization R of $\mathcal{S}_i$ satisfies the above two conditions, then R contains no inter-color dominances (by Lemma 39) and $\mathrm{sgn}(R) =$

(i,j), $\mathrm{sgn}(Z(R)) = (i',j')$ (note that $Z(R) = R \cap S_{i'}$). Therefore, the probability that R contains no inter-color dominances and $\mathrm{sgn}(R) = (i,j)$, $\mathrm{sgn}(Z(R)) = (i',j')$ is just the product $F(i',j') \cdot \pi^*_{i,j} \cdot \Pi_{i,j,i',j'}$, where $\pi^*_{i,j} = \pi(a_i) \cdot \pi(a_j)$ if $i \neq j$ and $\pi^*_{i,j} = \pi(a_i)$ if $i = j$, and $\Pi_{i,j,i',j'}$ is the product of all $(1-\pi(a_t))$ for $t \in \{i'+1, \ldots, i\}$ satisfying $\mathrm{cl}(a_t) \neq \mathrm{cl}(a_i)$ or $y(a_t) < y(a_j)$ or $y(a_{j'}) < y(a_t)$. Based on this, we can compute $F(i,j)$ as

$$F(i,j) = \sum_{(i',j') \in J_{i,j}} \left(F(i',j') \cdot \pi^*_{i,j} \cdot \Pi_{i,j,i',j'} \right) = \pi^*_{i,j} \cdot \sum_{(i',j') \in J_{i,j}} \left(F(i',j') \cdot \Pi_{i,j,i',j'} \right). \tag{4.1}$$

The straightforward way to compute each $F(i,j)$ takes $O(n^3)$ time, which results in an $O(n^5)$-time algorithm for computing $\Gamma_{\mathcal{S}}$.

Indeed, the runtime of the above algorithm can be drastically improved to $O(n^2 \log^2 n)$, by properly using dynamic 2D range trees with some tricks. Formally, we use a 2D range tree $\mathcal{T}$ built on a fixed collection of planar points and maintains the weights of these points. It supports the following three operations.

- Query$(\mathcal{T}, r)$: return the sum of weights of all the points in the query range r.
- Update$(\mathcal{T}, p, w)$: update the weight of point p to w.
- Multiply$(\mathcal{T}, r, \delta)$: multiply by a factor of δ the weight of every point in the range r. Note that this operation is reversible and the inverse of Multiply$(\mathcal{T}, r, \delta)$ is Multiply$(\mathcal{T}, r, 1/\delta)$.

We will show later that all of these operations can be done in $O(\log^2 n)$ time.

Two more notations are defined. For a legal pair (i,j), we use $(i,j)_{\searrow}$ (resp. $(i,j)_{\nwarrow}$) to represent the point $(x(a_i), y(a_j))$ (resp. $(x(a_j), y(a_i))$); see Figure 4.2. Also, let Quad(p) denote the northwest open quadrant of point p, i.e., $(-\infty, x(p)) \times (y(p), \infty)$. We give the complete solution in Algorithm 1 followed by the correctness analysis.

Correctness analysis. We compute $F(i,j)$ for each legal pair (i,j) by first enumerating i from 1 to n and then j in an order such points are visited from bottom to top; see the nested loop at Lines 8 and 14. For now, assume the fact, which we prove later, that the inner j-loop correctly computes $F(i,j)$ for all legal pairs (i,j) when i is fixed. We then have the following lemma.

Algorithm 1 Computing $\Gamma_{\mathcal{S}}$ in $O(n^2 \log^2 n)$ time.

1: **procedure** COMPUTE-$\Gamma_{\mathcal{S}}(\mathcal{S})$ ▷ Recall $\mathcal{S} = (S, \mathrm{cl}, \pi)$.
2: Sort all points in S such that $x(a_1) < \cdots < x(a_n)$.
3: Let $\mathcal{T}$ be the 2D range tree built on $\{(i,j)_{\searrow} : (i,j) \text{ is legal}\}$ with initial weights 0.
4: Let $\mathcal{T}_k$ be the 2D range tree built on $\{(i,j)_{\searrow} : (i,j) \text{ is legal and } \mathrm{cl}(a_i) = \mathrm{cl}(a_j) = k\}$ with initial weights 0, for every color k.
5: $prod = \prod_{i=1}^{n} (1 - \pi(a_i))$
6: $\Gamma_{\mathcal{S}} \leftarrow prod$
7: UPDATE$(\mathcal{T}, a_0, 1)$▷ This implies $F(0,0) = 1$. Also, no need to update $\mathcal{T}_{\mathrm{cl}(a_0)}$.
8: **for** $i \leftarrow 1$ **to** n **do**
9: $prod \leftarrow prod \cdot (1 - \pi(a_i))^{-1}$
10: $k \leftarrow \mathrm{cl}(a_i)$
11: MULTIPLY$(\mathcal{T}, \text{QUAD}(a_j), (1 - \pi(a_j))^{-1})$ for all $j \in \{1, \ldots, i\}$ such that $\mathrm{cl}(a_j) = k$.
12: MULTIPLY$(\mathcal{T}_k, \text{QUAD}(a_j), (1 - \pi(a_j))^{-1})$ for all $j \in \{1, \ldots, i\}$ such that $\mathrm{cl}(a_j) = k$.
13: Let $(\ell_1, \ldots, \ell_i)$ be a permutation of $(1, \ldots, i)$ such that $y(a_{\ell_1}) < \cdots < y(a_{\ell_i})$.
14: **for** $j \leftarrow \ell_1$ **to** ℓ_i **do**
15: **if** (i,j) is a legal pair **then** ▷ This implies that $\mathrm{cl}(a_j) = k$.
16: $F(i,j) \leftarrow \text{QUERY}(\mathcal{T}, \text{QUAD}((i,j)_{\nwarrow})) - \text{QUERY}(\mathcal{T}_k, \text{QUAD}((i,j)_{\nwarrow})$
17: $F(i,j) \leftarrow F(i,j) \cdot \pi^*_{i,j}$
18: $\Gamma_{\mathcal{S}} \leftarrow \Gamma_{\mathcal{S}} + F(i,j) \cdot prod$
19: MULTIPLY$(\mathcal{T}, \text{QUAD}(a_j), 1 - \pi(a_j))$
20: MULTIPLY$(\mathcal{T}_k, \text{QUAD}(a_j), 1 - \pi(a_j))$
21: **end if**
22: **end for**
23: Reverse all MULTIPLY operations executed in Lines 12, 19, 20.
24: UPDATE$(\mathcal{T}, (i,j)_{\searrow}, F(i,j))$ and UPDATE$(\mathcal{T}_k, (i,j)_{\searrow}, F(i,j))$ for every $j \in \{1, \ldots, i\}$ such that pair (i,j) is legal.
25: MULTIPLY$(\mathcal{T}, (-\infty, x(a_i)) \times \mathbb{R}, 1 - \pi(a_i))$
26: MULTIPLY$(\mathcal{T}_k, (-\infty, x(a_i)) \times \mathbb{R}, 1 - \pi(a_i))$
27: **end for**
28: **return** $\Gamma_{\mathcal{S}}$
29: **end procedure**

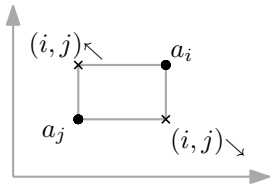


Figure 4.2: Illustrating $(i,j)_{\searrow}$ and $(i,j)_{\nwarrow}$ for a legal pair (i,j).

Lemma 40. *At the beginning of the i-th iteration of Line 8, the weight of $(i',j')_{\searrow}$ in $\mathcal{T}$, such that $i' < i$, is equal to $F(i',j') \cdot \prod_{p \in S \cap \|} (1 - \pi(p))$, where $\|$ denotes the open strip $(x(a_{i'}), x(a_i)) \times \mathbb{R}$. (See Figure 4.3(a).)*

Proof. This statement is trivially true for $i = 1$ as all the weights in $\mathcal{T}$ are equal to zero except that $F(0,0) = 1$. Assume the statement is true for the i-th iteration, we show it also holds for the $(i+1)$-th iteration. First, we can safely consider $\mathcal{T}$ unchanged throughout Lines 10-23 because although Lines 12 and 19 modify $\mathcal{T}$, these side-effects are reversed immediately in Line 23. After the inner j-loop is done, by our early assumption, we obtain the value of $F(i,j)$ for every legal pair (i,j) when i is fixed. These values are not currently stored in $\mathcal{T}$ but are needed for the next iteration. Thus, we update the weight of each $(i,j)_{\searrow} \in \mathcal{T}$ to $F(i,j)$, as stated in Line 24. We also need to multiply by the factor $(1 - \pi(a_i))$ the weight of each $(i',j')_{\searrow} \in \mathcal{T}$ that is to the left of a_i because a_i will be included in the strip as we proceed from i to $i+1$. This is handled by Line 25. As such, the statement is maintained for the $(i+1)$-th iteration, which completes the proof. □

With Lemma 40 in hand, we now give the proof of our aforementioned statement, as restated in Lemma 41.

Lemma 41. *Lines 16-17 correctly compute $F(i,j)$.*

Proof. Recall that $F(i,j) = \pi^*_{i,j} \cdot \sum_{(i',j') \in J_{i,j}} F(i',j') \cdot \Pi_{i,j,i',j'}$. By Lemma 40, at the beginning of the i-th round, the weight of each $(i',j')_{\searrow} \in \mathcal{T}$, where $i' < i$, is equal to $F(i',j') \cdot \prod_{p \in S \cap \|} (1 - \pi(p))$. This product is off from the ideal one, $F(i',j') \cdot \Pi_{i,j,i',j'}$, by a factor of $\prod_{p \in S^{(i)} \cap \square} (1 - \pi(p))$, where $S^{(i)} = \{p \in S : \mathrm{cl}(p) = \mathrm{cl}(a_i)\}$ and $\square$ denotes the box $(x(a_{i'}), x(a_i)) \times [y(a_j), y(a_{j'})]$; see Figure 4.3(b). To cancel this

factor, we observe that

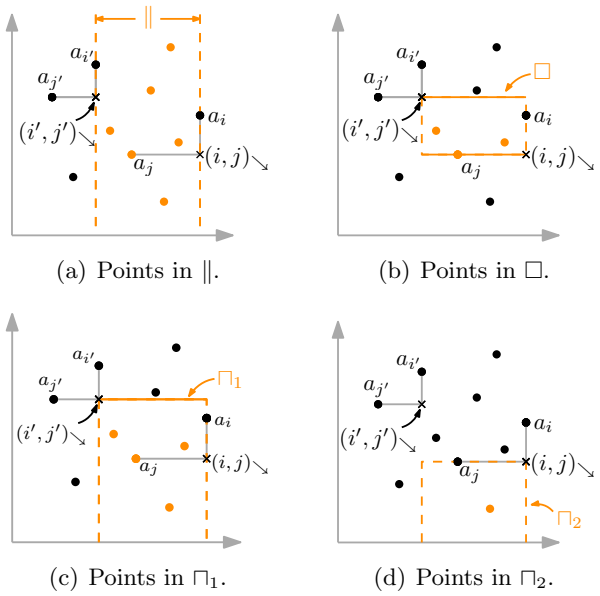
$$\prod_{p \in S^{(i)} \cap \square} (1 - \pi(p)) = \prod_{p \in S^{(i)} \cap \sqcap_1} (1 - \pi(p)) \Big/ \prod_{p \in S^{(i)} \cap \sqcap_2} (1 - \pi(p)),$$

where $\sqcap_1$ and $\sqcap_2$ respectively denote the three-sided rectangle $(x(a_{i'}), x(a_i)) \times (-\infty, y(a_{j'})]$ and $(x(a_{i'}), x(a_i)) \times (-\infty, y(a_j))$; see Figure 4.3(c) and 4.3(d). The former product ($\sqcap_1$) is canceled in Line 12, and the latter ($\sqcap_2$) is gradually accumulated back via $(j-1)$ calls of Line 19 as $a_{\ell_1}, \ldots, a_{\ell_{j-1}}$ are all below a_{ℓ_j}. Thus, the weight of each $(i', j')_{\searrow} \in \mathcal{T}$ is equal to $F(i', j') \times \Pi_{i,j,i',j'}$ right before $F(i, j)$ gets evaluated. Finally, the range query in Line 16 sums up the weight of every $(i', j')_{\searrow} \in \mathcal{T}$ such that $(i', j') \in J_{i,j}$. (Note that the subtraction in Line 16 is needed because $\textsc{Query}(\mathcal{T}, \textsc{Quad}((i, j)_{\nwarrow}))$ also counts the probabilities of those legal pairs that have the same color as $\text{cl}(a_i)$.) Therefore, the value of $F(i, j)$ is correctly computed after Line 17. □

Both the above lemmas can directly apply to the $\mathcal{T}_k$'s as we always query/update $\mathcal{T}$ and the $\mathcal{T}_k$'s in the same way. Finally, all $F(i, j)$'s are computed and added up into $\Gamma_{\mathcal{S}}$, which completes the correctness proof of the entire algorithm.

The overall runtime of Algorithm 1 is $O(n^2 \log^2 n)$ since there are $O(n^2)$ range queries and updates, each of which takes $O(\log^2 n)$ time. The space occupied by $\mathcal{T}$, denoted by $|\mathcal{T}|$, is $O(n^2 \log n^2) = O(n^2 \log n)$ as there are $O(n^2)$ legal pairs. Similarly, let n_k be the number of points in color k, and then $\mathcal{T}_k$ costs $O(n_k^2 \log n_k)$ space. Assume there are K colors in total. We have $n_1 + \cdots + n_K = n$ and thus $|\mathcal{T}_1| + \cdots + |\mathcal{T}_K| = O(n^2 \log n)$. The overall space complexity is $O(|\mathcal{T}| + |\mathcal{T}_1| + \cdots + |\mathcal{T}_K|) = O(n^2 \log n)$.

Finally, we discuss how to implement the augmented 2D range tree $\mathcal{T}$ to dynamically support the three operations $\textsc{Query}, \textsc{Update}$, and $\textsc{Multiply}$ in $O(\log^2 m)$ time, where m is the input size, and hence in $O(\log^2 n)$ time. We first describe how to implement a dynamic 1D range tree, $\mathcal{T}_{\text{1D}}$, built on the y-coordinates of a set of planar points, P, to support the three operations, where the range used in $\textsc{Query}$ and $\textsc{Multiply}$ is a 1D interval. The leaves, sorted by increasing y-coordinates, of $\mathcal{T}_{\text{1D}}$ are points in P with initial weight equal to 0. In addition, in each internal node, u, we store two fields, $sum(u)$ and $mul(u)$, where the former is the sum of weights in the subtree rooted at u and the latter is the multiplication factor that needs to



(a) Points in $\|$. (b) Points in $\square$.

(c) Points in $\sqcap_1$. (d) Points in $\sqcap_2$.

Figure 4.3: Illustrating Lemma 41. The orange color is only used to highlight each range and does *not* represent the color of each point. Dashed (resp. solid) boundaries are exclusive (resp. inclusive).

be applied to all the nodes in the subtree. For simplicity we use the notion $sum(u)$ to denote the weight of u if it is a leaf. Also, we set $sum(u) = 0$ and $mul(u) = 1$ initially.

Given a query/update range, we first identify $O(\log m)$ canonical nodes, $\mathcal{C}$, of $\mathcal{T}_{1D}$ via a recursive down-phase traversal. We then aggregate or modify the data in each canonical node. Finally, we refine the fields of those nodes along the path from every canonical node up to the root, as the recursion gradually terminates.

In the down-phase, when a non-leaf node u is visited, we call the following PUSH method to revise $sum(u)$ based on $mul(u)$ and then push the factor further to its two children. In the up-phase, we apply the COMBINE method to each node to readjust the sum. Between the down and up phase, we perform one of the following three operations.

- Add up $sum(u)$ for every $u \in \mathcal{C}$ for QUERY$(\mathcal{T}_{1D}, p)$.
- Update $sum(u)$ to w for the *only* element $u \in \mathcal{C}$ for UPDATE$(\mathcal{T}_{1D}, p, w)$.
- Multiply $mul(u)$ by a factor of δ for every $u \in \mathcal{C}$ for MULTIPLY$(\mathcal{T}_{1D}, p, \delta)$.

Algorithm 2 Implementation details of PUSH and COMBINE.

```
1: procedure PUSH(u)                          ▷ Only called in the down-phase.
2:     sum(u) ← sum(u) · mul(u)
3:     if u is not a leaf then
4:         mul(lchild(u)) ← mul(lchild(u)) · mul(u)
5:         mul(rchild(u)) ← mul(rchild(u)) · mul(u)
6:     end if
7:     mul ← 1
8: end procedure
9: procedure COMBINE(u)        ▷ Only called in the up-phase and we must have
   mul(u) = 1.
10:    sum(u) ← sum(lchild(u)) + sum(rchild(u))
11: end procedure
```

Next, we build our 2D range tree, $\mathcal{T}$, on the x-coordinates of the given input. For each node $u \in \mathcal{T}$, we build the aforementioned 1D range tree w.r.t. the set of points in u. We also store at u a tag indicating the multiplication factor that needs to be applied to the 1D range tree stored at u as well as all u's descendants. Given a 2D range query, we do a down-phase traversal identifying $O(\log m)$ canonical nodes of $\mathcal{T}$. For each visited node u during the traversal, we should apply the multiplication tag to the 1D tree stored at u and push it further to u's two children. This takes $O(\log m)$ time. Then, for every canonical node u, we spend another $O(\log m)$ time querying the 1D range tree stored at u, as stated above. Therefore, all three operations can be done in $O(\log^2 m)$ time, and $\mathcal{T}$ occupies $O(m \log m)$ space.

Remark. One may notice that the implementation above contains a flaw for MULTIPLY$(\mathcal{T}, r, \delta)$ when $\delta = 0$ because the inverse of this operation does not exist as $1/0$ is undefined. We can overcome this issue by adding in each node a *zero-counter* and counting the number of zero factors separately. That is, if MULTIPLY multiplies a factor of zero, we increment the zero-counter of each canonical node instead of modifying *sum* and *mul* fields; if MULTIPLY divides a factor of zero, we decrement the corresponding zero-counters. Also, when a QUERY is triggered, we simply return zero for those canonical nodes whose zero-counter is positive. This solves the problem without increasing the runtime of all three operations.

With the above argument, we conclude the following.

Theorem 42. *The CSD problem for $d = 2$ can be solved in $O(n^2 \log^2 n)$ time.*

4.2.2 Hardness results in higher dimensions

In this section, we prove the #P-hardness of the CSD problem for $d \geq 3$. Indeed, our hardness result is even stronger in that it applies to restricted versions of the CSD problem. There are two specializations of the CSD problem: in one all data points have distinct colors, in the other data points are bichromatic. We want our hardness result to cover these two specializations. Towards this end, we need to introduce a notion called *color pattern*.

A *partition* of a positive integer p is defined as a multi-set Δ of positive integers whose summation is p. In a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$, the coloring cl naturally induces a partition of $n = |S|$ given by the multi-set $\{|\mathrm{cl}^{-1}(p)| > 0 : p \in \mathbb{N}\}$, which we denote by $\Delta(\mathcal{S})$. Let $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ be an infinite sequence where Δ_p is a partition of p. We say $\mathcal{P}$ is a *color pattern* if it is "polynomial-time uniform", i.e., one can compute Δ_p for any given p in time polynomial in p. In addition, $\mathcal{P}$ is said to be *balanced* if $p - \max \Delta_p = \Omega(p^c)$ for some constant $c > 0$ (here $\max \Delta_p$ denotes the maximum in the multi-set Δ_p). Then we define the *CSD problem with respect to a color pattern* $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ as the (standard) CSD problem with the restriction that the input dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ must satisfy $\Delta(\mathcal{S}) = \Delta_n$ where $n = |S|$.

Besides specializing the CSD problem using the color pattern, we may also make assumptions for the existence probabilities of the points. An important case is that all points have the same existence probability of $\frac{1}{2}$. In this case, each of the 2^n subsets of S occurs as a realization of $\mathcal{S}$ with the same probability 2^{-n}, and computing $\Lambda_{\mathcal{S}}$ (or $\Gamma_{\mathcal{S}}$) is equivalent to counting the subsets of S satisfying the desired properties.

Our hardness result is presented in the following theorem.

Theorem 43. *Let $\mathcal{P}$ be any balanced color pattern. Then the CSD problem with respect to $\mathcal{P}$ is #P-hard for $d \geq 3$. In addition, even if the existence probabilities of the points are all restricted to be $\frac{1}{2}$, the CSD problem with respect to $\mathcal{P}$ remains #P-hard for $d \geq 7$.*

Note that our result above implies the hardness of both the distinct-color and bichromatic specializations. The former can be seen via a balanced color pattern

$\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ with $\Delta_p = \{1, \dots, 1\}$ (i.e., a multi-set consisting of p 1's), while the latter can be seen via a balanced color pattern $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ with $\Delta_p = \{\frac{p}{2}, \frac{p}{2}\}$ for even p and $\Delta_p = \{\frac{p-1}{2}, \frac{p+1}{2}\}$ for odd p. The proof of Theorem 43 is nontrivial, so we break it into several stages.

4.2.2.1 Relation to counting independent sets

For a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$, define $G_{\mathcal{S}} = (S, E_{\mathcal{S}})$ as the (undirected) graph with vertex set S and edge set $E_{\mathcal{S}} = \{(a, b) : a, b \in S \text{ with } \mathrm{cl}(a) \neq \mathrm{cl}(b) \text{ and } a \succ b\}$. Since the edges of $G_{\mathcal{S}}$ correspond one-to-one to the inter-color dominances in S, it is clear that a subset $A \subseteq S$ contains no inter-color dominances iff A corresponds to an independent set of $G_{\mathcal{S}}$. If $\pi(a) = \frac{1}{2}$ for all $a \in S$, then we immediately have the equation $\Gamma_{\mathcal{S}} = Ind(G_{\mathcal{S}})/2^n$, where $Ind(G_{\mathcal{S}})$ is the number of the independent sets of $G_{\mathcal{S}}$. This observation intuitively tells us the hardness of the CSD problem, as independent-set counting is a well-known #P-complete problem. Although we are still far away from proving Theorem 43 (because for a given graph G it is not clear how to construct a colored stochastic dataset $\mathcal{S}$ such that $G_{\mathcal{S}} \cong G$, i.e., $G_{\mathcal{S}}$ is isomorphic to G), it is already clear that we should reduce from some independent-set-counting problem. Regarding independent-set counting, the strongest known result is the following theorem obtained by Xia et al. [35], which will be used as the origin of our reduction.

Theorem 44. *Counting independent sets for 3-regular planar bigraphs is #P-complete.*

For a graph $G = (V, E)$, we say a map $f : V \to \mathbb{R}^d$ is a *dominance-preserving embedding* (DPE) of G to $\mathbb{R}^d$ if it satisfies the condition that $(u, v) \in E$ iff $f(u) \succ f(v)$ or $f(v) \succ f(u)$. We define the *dimension*, $\dim(G)$, of G as the smallest number d such that there exists a DPE of G to $\mathbb{R}^d$ (if such a number does not exist, we say G is of infinite dimension). We have seen above the relation between independent-set counting and the CSD problem with existence probabilities equal to $\frac{1}{2}$. Interestingly, with general existence probabilities, the CSD problem can be related to a much stronger version of independent-set counting, which we call *cardinality-sensitive independent-set counting*.

Definition 45. *Let c be a fixed integer. The c-cardinality-sensitive independent-set counting (c-CSISC) problem is defined as follows. The input consists of a graph $G = (V, E)$ and a c-tuple $\Phi = (V_1, \ldots, V_c)$ of disjoint subsets of V. The task of the problem is to output, for every c-tuple $(n_1, \ldots, n_c)$ of integers where $0 \leq n_i \leq |V_i|$, the number of the independent sets $I \subseteq V$ of G satisfying $|I \cap V_i| = n_i$ for all $i \in \{1, \ldots, c\}$. We denote the desired output by $Ind_\Phi(G)$, which can be represented by a sequence of $\prod_{i=1}^{c}(|V_i| + 1)$ integers. Note that the 0-CSISC problem is just the conventional independent-set counting.*

Lemma 46. *Given any graph $G = (V, E)$ with a DPE $f : V \to \mathbb{R}^d$ and a c-tuple $\Phi = (V_1, \ldots, V_c)$ of disjoint subsets of V, one can construct in polynomial time a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ in $\mathbb{R}^d$, with* cl *injective, such that* (1) *$G_\mathcal{S} \cong G$ and* (2) *$Ind_\Phi(G)$ can be computed in polynomial time if $\Gamma_\mathcal{S}$ is provided. In particular, the c-CSISC problem for a class $\mathcal{G}$ of graphs is polynomial-time reducible to the CSD problem in $\mathbb{R}^d$, given an oracle that computes for any graph in $\mathcal{G}$ a DPE of that graph to $\mathbb{R}^d$.*

Proof. Suppose $|V| = \{v_1, \ldots, v_n\}$. We construct the colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ as follows. Define $S = \{a_1, \ldots, a_n\}$ where $a_i = f(v_i) \in \mathbb{R}^d$ and set $\mathrm{cl}(a_i) = i$ (so cl is injective). Let $S_1, \ldots, S_c$ be the (disjoint) subsets of S corresponding to $V_1, \ldots, V_c$ respectively, i.e., $S_i = \{a_j : v_j \in V_i\}$. Without loss of generality, we may assume $S_1, \ldots, S_c$ are all nonempty. For all points $a \in S_i$, we define $\pi(a) = 4^{-n^{c-i+1}}$ (note that this real number can be represented in polynomial length). Then for all points $a \in S \backslash (\bigcup_{i=1}^{c} S_i)$, we define $\pi(a) = \frac{1}{2}$. With $\mathcal{S}$ constructed above, we already have $G_\mathcal{S} \cong G$, since f is a DPE and all the points in S have distinct colors. It suffices to show how to "recover" $Ind_\Phi(G)$ from $\Gamma_\mathcal{S}$. Equivalently, we have to compute, for every c-tuple $\phi = (n_1, \ldots, n_c)$ of integers where $0 \leq n_i \leq |S_i|$, the number of the subsets $A \subseteq S$ containing no inter-color dominances and satisfying $|A \cap S_i| = n_i$ for all $i \in \{1, \ldots, c\}$ (we use $\mathcal{A}_\phi$ to denote the collection of these subsets). For each c-tuple $\phi = (n_1, \ldots, n_c)$ with $0 \leq n_i \leq |S_i|$, we notice that any $A \in \mathcal{A}_\phi$ occurs as a realization of $\mathcal{S}$ with probability

$$P_\phi = \frac{1}{2^{n-m}} \prod_{i=1}^{c} \left(\frac{1}{4^{n^{c-i+1}}}\right)^{n_i} \left(1 - \frac{1}{4^{n^{c-i+1}}}\right)^{|S_i|-n_i},$$

where $m = \sum_{i=1}^{c} |S_i|$. By setting $N = \prod_{i=1}^{c}(|S_i| + 1)$, we have in total N c-tuples $\phi_1, \ldots, \phi_N$ (of integers) to be considered (N is polynomial in n as c is constant). Suppose $\phi_1, \ldots, \phi_N$ are already sorted in lexicographical order from small to large. Our first key observation is that $P_{\phi_i} > 2^n P_{\phi_{i+1}}$ for all $i \in \{1, \ldots, N-1\}$. To see this, assume $\phi_i = (n_1, \ldots, n_c)$ and $\phi_{i+1} = (n'_1, \ldots, n'_c)$. Note that $\phi_1, \ldots, \phi_N$ are sorted in lexicographical order, so there exists $k \in \{1, \ldots, c\}$ such that $n_j = n'_j$ for all $j < k$ and $n'_k = n_k + 1$. Then it is easy to see that

$$\frac{P_{\phi_i}}{P_{\phi_{i+1}}} \geq \frac{1 - 4^{-n^{c-k+1}}}{4^{-n^{c-k+1}}} \prod_{j=k+1}^{c} (4^{-n^{c-j+1}})^{|S_j|}.$$

If $k = c$, we already have $P_{\phi_i} > 2^n P_{\phi_{i+1}}$. For the case of $k < c$, since $\sum_{j=k+1}^{c} |S_j| \leq n - 1$, the above inequality implies that

$$\frac{P_{\phi_i}}{P_{\phi_{i+1}}} \geq \frac{(1 - 4^{-n^{c-k+1}}) \cdot 4^{-(n-1)\cdot n^{c-k}}}{4^{-n^{c-k+1}}} > 2^n.$$

With this observation in hand, we now consider how to compute $|\mathcal{A}_{\phi_i}|$ for all $i \in \{1, \ldots, N\}$ from $\mathit{\Gamma}_{\mathcal{S}}$. It is clear that

$$\mathit{\Gamma}_{\mathcal{S}} = \sum_{i=1}^{N} P_{\phi_i} \cdot |\mathcal{A}_{\phi_i}|.$$

For $j \in \{1, \ldots, N\}$, we set $\gamma_j = \sum_{i=j+1}^{N} P_{\phi_i} \cdot |\mathcal{A}_{\phi_i}|$. By the facts that $P_{\phi_i} > 2^n P_{\phi_{i+1}}$ and $\sum_{i=1}^{N} |\mathcal{A}_{\phi_i}| \leq 2^n$, we can deduce $P_{\phi_i} > \gamma_i$ for all $i \in \{1, \ldots, N\}$. Then we are ready to compute $|\mathcal{A}_{\phi_1}|, \ldots, |\mathcal{A}_{\phi_N}|$ in order. Since $P_{\phi_1} > \gamma_1$, $|\mathcal{A}_{\phi_1}|$ must be the greatest integer that is smaller than or equal to $\mathit{\Gamma}_{\mathcal{S}}/P_{\phi_1}$, and hence can be immediately computed. Suppose now $|\mathcal{A}_{\phi_1}|, \ldots, |\mathcal{A}_{\phi_{m-1}}|$ are already computed, and we consider $|\mathcal{A}_{\phi_m}|$. Via $|\mathcal{A}_{\phi_1}|, \ldots, |\mathcal{A}_{\phi_{m-1}}|$ and $\mathit{\Gamma}_{\mathcal{S}}$, we may compute γ_{m-1}. Because $P_{\phi_m} \geq \gamma_m$, $|\mathcal{A}_{\phi_m}|$ must be the greatest integer that is smaller than or equal to γ_{m-1}/P_{ϕ_m}, and hence can be computed directly. In this way, we are able to compute all $|\mathcal{A}_{\phi_1}|, \ldots, |\mathcal{A}_{\phi_N}|$ and equivalently $\mathit{Ind}_{\mathit{\Phi}}(G)$ (in polynomial time). The statements in the lemma follow readily. □

Another ingredient to be used in the proof of Theorem 43 is a lemma regarding color patterns.

Lemma 47. *Let $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ be a balanced color pattern. Given a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ in $\mathbb{R}^d$, with* cl *injective, if $G_{\mathcal{S}}$ is a bipartite graph, then one can construct in polynomial time another colored stochastic dataset $\mathcal{S}' = (S', \mathrm{cl}', \pi')$ in $\mathbb{R}^d$ satisfying* (1) $\Gamma_{\mathcal{S}'} = \Gamma_{\mathcal{S}}$, (2) $S \subseteq S'$, (3) $\pi'(a) = \frac{1}{2}$ *for any* $a \in S' \backslash S$, (4) $\langle \mathcal{S}' \rangle$ *is an instance of the CSD problem with respect to* $\mathcal{P}$.

Proof. Since $\mathcal{P}$ is balanced, we can find an constants $c > 0$ such that $n - \max \Delta_n \geq n^c$ for any sufficiently large n. Suppose $G_{\mathcal{S}} = (V \cup V', E)$ where $|V| = n$ and $|V'| = n'$. We may write $S = \{a_1, \dots, a_{n+n'}\}$ where $a_1, \dots, a_n$ correspond to the vertices in V and $a_{n+1}, \dots, a_{n+n'}$ correspond to those in V'. Because cl is injective (i.e., the points in S are of distinct colors), we have that $a_1, \dots, a_n$ do not dominate each other, and the same holds for $a_{n+1}, \dots, a_{n+n'}$. Set $N = \max\{2n + n', (n')^{1/c}\}$. Now we construct $\mathcal{S}' = (S', \mathrm{cl}', \pi')$ as follows. First, we pick a set A of $N - (n + n')$ points in $\mathbb{R}^d$ which do not dominate each other and do not form dominances with any points in S. Set $S' = S \cup A$, so $S \subseteq S'$ and $|S'| = N$. The points in A are used as dummy points, and can never influences $\Gamma_{\mathcal{S}'}$ (since they are not involved in any dominances). With a slight abuse of notation, we also use $a_1, \dots, a_{n+n'}$ to denote the non-dummy points in S'. We then define π' as $\pi'(a) = \pi(a)$ for $a \in S$ and $\pi'(a) = \frac{1}{2}$ for $a \in A$. It suffices to assign colors to the points in S', i.e., define the coloring function cl$'$. Since we want $\langle \mathcal{S}' \rangle$ to be an instance of the CSD problem with respect to $\mathcal{P}$, the coloring cl$'$ must induce the partition Δ_N of N. Suppose $\Delta_N = \{r_1, \dots, r_k\}$ (as a multi-set) where $r_1 \geq \cdots \geq r_k$. Let l be the smallest integer such that $\sum_{i=1}^{l} r_i \geq n$. It is easy to see that $\sum_{i=l+1}^{k} r_i \geq n'$. Indeed, if $l = 1$, then we have

$$\sum_{i=2}^{m} r_i = N - \max \Delta_N \geq N^c \geq n'$$

by assumption. In the case of $l > 1$, we have that $\sum_{i=1}^{l} r_i < 2n$ and thus $\sum_{i=l+1}^{k} r_i > N - 2n \geq n'$. This fact implies that we are able to define the coloring function cl$'$ with image $\{1, \dots, k\}$ such that (1) there are exactly r_i points in S' mapped to the color i by cl$'$, (2) $\mathrm{cl}'(a) \in \{1, \dots, l\}$ for any $a \in \{a_1, \dots, a_n\}$, (3) $\mathrm{cl}'(a) \in \{l + 1, \dots, m\}$ for any $a \in \{a_{n+1}, \dots, a_{n+n'}\}$. With this cl$'$, we have that $\mathrm{cl}'(a_i) \neq \mathrm{cl}'(a_j)$ for any $i \in \{1, \dots, n\}$ and $j \in \{n + 1, \dots, n + n'\}$. Therefore, if two points $a_i, a_j \in S$ form an inter-color dominance in with respect to cl, then they also form an inter-color dominance with respect to cl$'$, and vice versa. Since the dummy points in A can

never contribute inter-color dominances, we have $\Gamma_{\mathcal{S}'} = \Gamma_{\mathcal{S}}$, which completes the proof. □

4.2.2.2 #P-hardness for $d \geq 3$

In this section, we prove the first statement of Theorem 43, by providing a reduction from the independent-set counting problem for 3-regular planar bipartite graphs. Let $G = (V \cup V', E)$ be a 3-regular planar bipartite graph. Suppose $|V| = |V'| = n$ (note that we must have $|V| = |V'|$ for G is 3-regular); then $|E| = 3n$. Instead of working on G directly, we shall first construct a new graph G^* based on G, and try to embed G^* into $\mathbb{R}^3$. Set $\lambda = 100n^2$. We define G^* as the graph obtained from G by inserting 2λ new vertices into each edge of G, i.e., replacing each edge of G with a chain of 2λ new vertices (see Figure 4.4). With an abuse of notation, V and V' are also used to denote the corresponding subsets of the vertices of G^*. Note that G^* is also bipartite, in which V and V' belong to different parts. We use U (resp., U') to denote the set of the inserted vertices of G^* which belong to the same part as V (resp., V'). Then the two parts of G^* are $V \cup U$ and $V' \cup U'$. For each edge $e \in E$ of G, we denote by U_e (resp. U'_e) the set of the λ vertices in U (resp., U') which are inserted into the edge e.

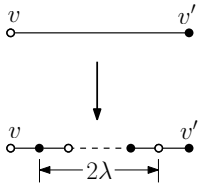


Figure 4.4: Inserting new vertices into each edge of G.

It is not surprising that the independent sets of G are strongly related to those of G^*. Indeed, as we will show, counting independent sets for G can be done by solving the 4-CSISC instance $\langle G^*, (V, V', U, U')\rangle$. Define $Ind_{p,p'}$ as the number of the independent sets I of G such that $|I \cap V| = p$, $|I \cap V'| = q$. Also, define $Ind^*_{p,p',q,q'}$ as the number of the independent sets I^* of G^* such that $|I^* \cap V| = p$, $|I^* \cap V'| = p'$, $|I^* \cap U| = q$, $|I^* \cap U'| = q'$.

Lemma 48. *For any $p, p' \in \{0, \ldots, n\}$, we have $Ind_{p,p'} = Ind^*_{p,p',3\lambda p,3\lambda n-3\lambda p}$. In*

particular,

$$Ind(G) = \sum_{i=0}^{n}\sum_{j=0}^{n} Ind_{i,j} = \sum_{i=0}^{n}\sum_{j=0}^{n} Ind^*_{i,j,3\lambda i,3\lambda n-3\lambda i}.$$

Proof. Fixing $p, p' \in \{0, \ldots, n\}$, we denote by $\mathcal{I}$ the collection of the independent sets I of G such that $|I \cap V| = p$, $|I \cap V'| = p'$. Also, we denote by $\mathcal{I}^*$ the collection of the independent sets I^* of G^* such that $|I^* \cap V| = p$, $|I^* \cap V'| = p'$, $|I^* \cap U| = 3\lambda p$, $|I^* \cap U'| = 3\lambda n - 3\lambda p$. It suffices to establish a one-to-one correspondence between $\mathcal{I}$ and $\mathcal{I}^*$.

Let $I \in \mathcal{I}$ be an element. If $e = (v, v') \in E$ is an edge of G (where $v \in V$ and $v' \in V'$), we say e is of Type-1 if $v \in I$ (and hence $v' \notin I$), otherwise of Type-2. Recall that for each $e \in E$, U_e (resp., U'_e) denotes the set of the λ vertices in U (resp., U') which are inserted to the edge e. Now let I^* be the set consists of the vertices in I, the vertices in U_e for all Type-1 edges e, and the vertices in U'_e for all Type-2 edges e. Clearly, I^* is an independent set of G^*. Furthermore, by the definition of $\mathcal{I}$ and the fact that G is 3-regular, we know that G has $3p$ Type-1 edges and $3n-3p$ Type-2 edges. It follows that $|I^* \cap V| = p$, $|I^* \cap V'| = p'$, $|I^* \cap U| = 3\lambda p$, $|I^* \cap U'| = 3\lambda n - 3\lambda p$. Thus, $I^* \in \mathcal{I}^*$. By mapping I to I^*, we obtain a map from $\mathcal{I}$ to $\mathcal{I}^*$, which is obviously injective.

To see it is surjective, let $I^* \in \mathcal{I}^*$ be an element. Set $I = I^* \cap (V \cup V')$. We claim that $I \in \mathcal{I}$ and I is mapped to I^* by our map defined above. First, since I^* is an independent set of G^*, we must have $|I^* \cap (U_e \cup U'_e)| \leq \lambda$ for any edge $e = (v, v') \in E$ of G (with equality only if at least one of v and v' is in I). But $|I^* \cap (U \cup U')| = 3\lambda n = \lambda|E|$, which implies $|I^* \cap (U_e \cup U'_e)| = \lambda$ for all $e \in E$. It follows that for every edge $e = (v, v') \in E$, v and v' are not included in I simultaneously, i.e., I is an independent set of G. In addition, $|I \cap V| = |I^* \cap V| = p$, $|I \cap V'| = |I^* \cap V'| = p'$. Therefore, $I \in \mathcal{I}$. To see I is mapped to I^*, we apply again the fact that $|I^* \cap (U_e \cup U'_e)| = \lambda$ for any $e \in E$. Based on this, we further observe that for any $e \in E$, either $U_e \subseteq I^*$ or $U'_e \subseteq I^*$ (since I^* is an independent set of G^*). As before, we say an edge $e = (v, v') \in E$ (with $v \in V$ and $v' \in V'$) is of Type-1 if $v \in I$, otherwise of Type-2. Note that if an edge $e \in E$ is of Type-1, we must have $U_e \subseteq I^*$ (and then $I^* \cap U'_e = \emptyset$). Since G has $3p$ Type-1 edges, $|I^* \cap U| \geq 3\lambda p$. But in fact $|I^* \cap U| = 3\lambda p$ as $I^* \in \mathcal{I}^*$. So the only possibility is that $U_e \subseteq I^*$ (and $I^* \cap U'_e = \emptyset$) for all Type-1 edges e and $U'_e \subseteq I^*$ (and $I^* \cap U_e = \emptyset$) for all Type-2

edges e. As a result, I is mapped to I^* and $|\mathcal{I}| = |\mathcal{I}^*|$, completing the proof. □

Now it suffices to reduce the 4-CSISC instance $\langle G^*, (V, V', U, U')\rangle$ to an instance $\langle \mathcal{S}\rangle$ of the CSD problem in $\mathbb{R}^3$ with respect to a given balanced color pattern $\mathcal{P}$. Due to Lemmas 46 and 47, the only thing we need for the reduction is a DPE of G^* to $\mathbb{R}^3$. Therefore, our next step is to show $\dim(G^*) \leq 3$ and construct explicitly a DPE of G^* to $\mathbb{R}^3$ (in polynomial time), which is the most non-obvious part of the proof.

Recall that the two parts of G^* are $V \cup U$ and $V' \cup U'$. The DPE that we are going to construct makes the image of each vertex in $V' \cup U'$ dominate the images of its adjacent vertices in $V \cup U$. We first consider the embedding for the part $V \cup U$. Our basic idea is to map the vertices in $V \cup U$ to the plane $H : x + y + z = 0$ in $\mathbb{R}^3$. Note that by doing this we automatically prevent their images from dominating each other. However, the locations of (the images of) these vertices on H should be carefully chosen so that later we are able to further embed the part $V' \cup U'$ (to $\mathbb{R}^3$) to form a DPE. Basically, we map $V \cup U$ to H through two steps. In the first step, the vertices in $V \cup U$ are mapped to $\mathbb{R}^2$ via a map $\varphi : V \cup U \to \mathbb{R}^2$ to be constructed. Then in the second step, we properly project $\mathbb{R}^2$ onto H via another map $\psi : \mathbb{R}^2 \to H$. By composing ψ and φ, we obtain the desired map $\psi \circ \varphi : V \cup U \to H$, which gives us the embedding for $V \cup U$.

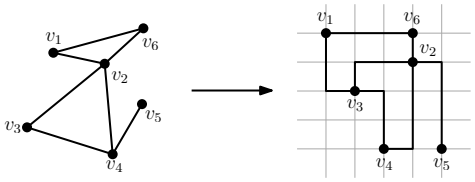


Figure 4.5: An orthogonal grid drawing.

To construct φ, we need a notion about graph drawing. Let $K = (\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Z}) \subset \mathbb{R}^2$ be the grid. An *orthogonal grid drawing* (OGD) of a (planar) graph is a planar drawing with image in the grid K such that the vertices are mapped to the grid points $\mathbb{Z}^2$. Note that an OGD draws the edges of the graph as (non-intersecting) orthogonal curves in $\mathbb{R}^2$ consisting of unit-length horizontal/vertical segments each of which connects two adjacent grid points (see Figure 4.5). We will apply the following result from [36].

Theorem 49. *For any t-vertex planar graph of (maximum) degree 3, one can compute in polynomial time an OGD with image in $K \cap Q_{3t}$ where Q_i denotes the square $[1, i]^2 \subset \mathbb{R}^2$.*

Consider the original 3-regular planar bipartite graph $G = (V \cup V', E)$. By applying the above theorem, we can find an OGD g for G with image in $K \cap Q_{6n}$. For each vertex $v \in V \cup V'$ of G, we denote by $g(v)$ the image of v in $\mathbb{R}^2$ under the OGD g. Also, for each edge $e = (v, v') \in E$ of G, we denote by $g(e)$ the image of e under g, which is an orthogonal curve in $\mathbb{R}^2$ connecting $g(v)$ and $g(v')$. With the OGD g in hand, we construct the map φ as follows.

For all $v \in V$, we simply define $\varphi(v) = g(v)$. To determine $\varphi(u)$ for $u \in U$, we consider the vertices in U_e for each edge $e \in E$ of G separately. Suppose $e = (v, v')$ and $U_e = \{u_1, \ldots, u_\lambda\}$ where $u_1, \ldots, u_\lambda$ are sorted in the order they appear on e (from v to v'). Consider the curve $g(e)$. Since g is an OGD, $g(e)$ must consist of unit-length horizontal/vertical segments (each of which connects two grid points). The total number m of these unit segments is upper bounded by $(6n)^2$ as $g(e) \subset K \cap Q_{6n}$. Now we pick a set P_e of λ (distinct) points on $g(e)$ as follows.

• The $m-1$ grid points in the interior of $g(e)$ are included in P_e (see Figure 4.6(a)).

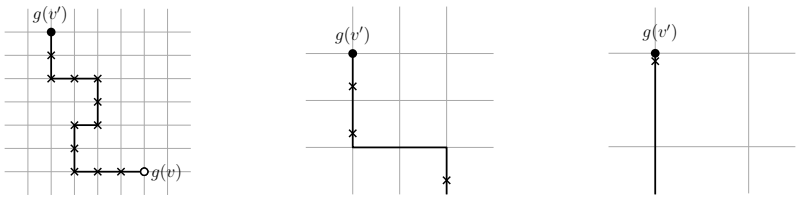
• On each unit vertical segment of $g(e)$, we pick the point with distance 0.3 from the bottom endpoint and include it to P_e (see Figure 4.6(b)).

• On the unit segment of $g(e)$ adjacent to $g(v')$, we pick the point with distance 0.01 from $g(v')$ and include it to P_e (see Figure 4.6(c)).

• Note that the number of the above three types of points is at most $2m \leq 72n^2 < \lambda$. To make $|P_e| = \lambda$, we then arbitrarily pick more (distinct) points on $g(e)$ which have distances at least 0.4 to any grid point, and add them to P_e.

Suppose $P_e = \{r_1, \ldots, r_\lambda\}$ where $r_1, \ldots, r_\lambda$ are sorted in the order they appear on the curve $g(e)$ (from $g(v)$ to $g(v')$). We then define $\varphi(u_i) = r_i$. We do the same thing for every edge $e \in E$ of G. In this way, we determine $\varphi(u)$ for all $u \in U$ and complete defining the map φ.

The next step, as mentioned before, is to project $\mathbb{R}^2$ onto H. The projection map $\psi : \mathbb{R}^2 \to H$ is defined as $\psi : (x, y) \mapsto (x+y, y-x, -2y)$. Then the composition $\psi \circ \varphi : V \cup U \to H$ gives us the first part of our DPE. The remaining task is to embed the part $V' \cup U'$ to $\mathbb{R}^3$, which completes the construction of our DPE. We



(a) (b) (c)

Figure 4.6: The construction of P_e.

must guarantee that the image of each vertex $w' \in V' \cup U'$ dominates and only dominates the images of the vertices in $V \cup U$ adjacent to w'. To achieve this, we first establish an important property of the map $\psi \circ \varphi : V \cup U \to H$ constructed above. For a finite set A of points in $\mathbb{R}^d$, we define a point c-max$(A) \in \mathbb{R}^d$ as the *coordinate-wise* maximum of A, i.e., the i-th coordinate of c-max(A) is the maximum of the i-th coordinates of all points in A, for all $i \in \{1, \ldots, d\}$.

Lemma 50. *For each vertex $w' \in V' \cup U'$, let $\mathrm{Adj}_{w'} \subseteq V \cup U$ be the set of the vertices adjacent to w' in G^*, and $A_{w'} = (\psi \circ \varphi)(\mathrm{Adj}_{w'}) \subset \mathbb{R}^3$ be the set of the corresponding images under $\psi \circ \varphi$. Then for any $w \in V \cup U$ and $w' \in V' \cup U'$, the point c-max$(A_{w'}) \in \mathbb{R}^3$ dominates $(\psi \circ \varphi)(w)$ iff $w \in \mathrm{Adj}_{w'}$.*

Proof. The "if" part is obvious, because c-max(A) clearly dominates every point in A for any (finite) $A \subset \mathbb{R}^d$ with $|A| \geq 2$ (note that $|A_{w'}| \geq 2$ for any $w' \in V' \cup U'$). It suffices to prove the "only if" part. For a point $p \in \mathbb{R}^3$, we denote by H_p the set of the points on the plane H which are dominated by p. We first observe that if $H_p \neq \emptyset$, then the preimage $\psi^{-1}(H_p)$ of H_p under ψ (which is a region in $\mathbb{R}^2$) must be a (closed) right-angled isosceles triangle in $\mathbb{R}^2$ whose hypotenuse is horizontal (we call these kinds of triangles *standard* triangles). To see this, assume $p = (x_p, y_p, z_p)$ and $H_p \neq \emptyset$ (this is equivalent to saying $x_p + y_p + z_p > 0$). Then $\psi^{-1}(H_p)$ consists of all the points $(x, y) \in \mathbb{R}^2$ satisfying $x + y \leq x_p$, $y - x \leq y_p$, $y \geq -z_p/2$, and hence is a standard triangle. Furthermore, it is easy to see that if $p =$ c-max(A) for a finite set $A \subset H$ with $|A| \geq 2$, then $H_p \neq \emptyset$ and $\psi^{-1}(H_p)$ is the minimal standard triangle containing $\psi^{-1}(A)$ (by "minimal" we mean that any standard triangle containing $\psi^{-1}(A)$ is a superset of $\psi^{-1}(H_p)$, both as subsets of $\mathbb{R}^2$, see Figure 4.7). Therefore, we only need to show that for any vertex $w' \in V' \cup U'$, the

minimal standard triangle containing $\psi^{-1}(A_{w'}) = \varphi(\mathrm{Adj}_{w'})$ does not contain $\varphi(w)$ for any vertex $w \in (V \cup U) \backslash \mathrm{Adj}_{w'}$. We consider two cases, $w' \in V'$ and $w' \in U'$.

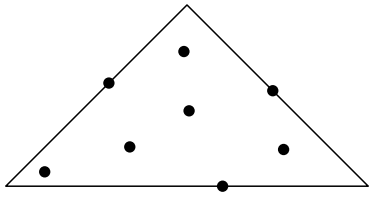


Figure 4.7: The minimal standard triangle in $\mathbb{R}^2$ containing a set of points.

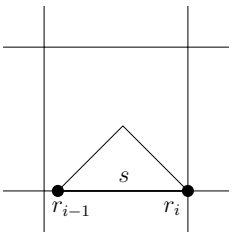


Figure 4.8: The case that s is horizontal.

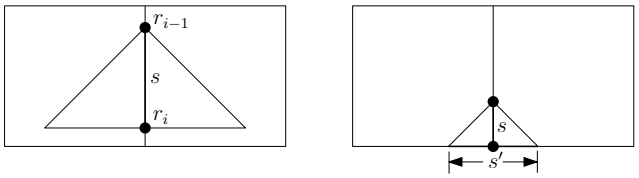


Figure 4.9: The case that s is vertical.

In the case of $w' \in V'$, $\mathrm{Adj}_{w'}$ consists of three vertices (for G is 3-regular) in U, say w_1, w_2, w_3. Recall that g is the OGD of G used in constructing the map φ. By recalling our construction of φ, we see that each of $\varphi(w_1), \varphi(w_2), \varphi(w_3)$ has distance 0.01 from $g(w')$. On the other hand, one can easily verify that for any vertex $w \in (V \cup U) \backslash \mathrm{Adj}_{w'}$, $\varphi(w)$ is "far away" from $g(w')$ (more precisely, with distance at least 0.3). Therefore, the minimal standard triangle containing $\varphi(w_1), \varphi(w_2), \varphi(w_3)$ does not contain $\varphi(w)$ for any vertex $w \in (V \cup U) \backslash \mathrm{Adj}_{w'}$.

In the case of $w' \in U'$, we may assume $w' \in U'_e$ for some edge $e = (v, v') \in E$ of G. Then $\mathrm{Adj}_{w'}$ consists of two vertices in $\{v\} \cup U_e$, say w_1, w_2. Recall that P_e is the

set of the λ points chosen on the curve $g(e)$ for sake of defining $\varphi(u)$ for $u \in U_e$. As before, we suppose $P_e = \{r_1, \ldots, r_\lambda\}$ where $r_1, \ldots, r_\lambda$ are sorted in the order they appear on the curve $g(e)$ (from $g(v)$ to $g(v')$). For convenience, set $r_0 = g(v)$. Then we may assume $\varphi(w_1) = r_{i-1}$ and $\varphi(w_2) = r_i$ for some $i \in \{1, \ldots, \lambda\}$. Let $s = \overline{r_{i-1}r_i}$ be the segment in $\mathbb{R}^2$ with endpoints r_{i-1} and r_i, and $\triangle$ be the minimal standard triangle containing r_{i-1} and r_i. Since all the grid points on $g(e)$ are included in P_e, s must be a horizontal or vertical segment contained in $g(e)$. Furthermore, the interior of s does not contain $\varphi(w)$ for any vertex $w \in V \cup U$ and in particular does not contain any grid points. We discuss two cases separately: s is horizontal and s is vertical. Recall that $K = (\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times \mathbb{Z}) \subset \mathbb{R}^2$ is the grid.

If s is horizontal, then $\triangle$ is just the standard triangle having s as its hypotenuse (see Figure 4.8). In this case, we have $\triangle \cap K = s$, which implies that $\triangle$ does not contain $\varphi(w)$ for any vertex $w \in (V \cup U) \backslash \{w_1, w_2\}$.

For the case that s is vertical, assume that r_{i-1} is the top endpoint and r_i is the bottom one. Then r_{i-1} is the right-angled vertex of $\triangle$, and r_i is the midpoint of the hypotenuse of $\triangle$. If r_i is not a grid point, we again have $\triangle \cap K = s$ and thus we are done (see the left part of Figure 4.9). If r_i is a grid point, the distance between r_{i-1} and r_i must be 0.3, by our construction of P_e. In this situation, $\triangle \cap K$ consists of s and a horizontal segment s' of length 0.6 which is the hypotenuse of $\triangle$ (see the right part of Figure 4.9). We claim that $\varphi(w)$ is not on s' for any vertex $w \in (V \cup U) \backslash \{w_2\}$. Indeed, by our construction of φ, if $\varphi(w)$ is in the interior of some unit horizontal segment, then $\varphi(w)$ is either with distance 0.01 from $g(v')$ for some $v' \in V'$ or with distance at least 0.4 from any grid point.

Thus, in each of the cases, $\varphi(w)$ is "far away" from r_i (more precisely, with distance at least 0.4). But any point on s' has distance at most 0.3 from r_i. Therefore, $\varphi(w)$ is not on s'. It immediately follows that $\triangle$ does not contain $\varphi(w)$ for any vertex $w \in (V \cup U) \backslash \{w_1, w_2\}$, which completes the proof. $\square$

Once the above property is revealed, the construction of the map $V' \cup U' \to \mathbb{R}^3$ is quite simple: we just map each vertex $w' \in V' \cup U'$ to the point c-max$(A_{w'}) \in \mathbb{R}^3$. Now we complete constructing the embedding of G^* to $\mathbb{R}^3$, and need to verify it is truly a DPE. Lemma 50 already guarantees that the image of each $w' \in V' \cup U'$ dominates (the images of) the vertices in $\mathrm{Adj}_{w'}$ (i.e., the vertices in $V \cup U$ that are

adjacent to w') but does not dominate (the images of) any other vertices in $V \cup U$. So it suffices to show that the images of the vertices in $V' \cup U'$ do not dominate each other. Let $w'_1, w'_2 \in V' \cup U'$ be two distinct vertices, and assume that c-max$(A_{w'_1})$ dominates c-max$(A_{w'_2})$. Then we must have c-max$(A_{w'_1})$ dominates the points in $A_{w'_2}$. By Lemma 50, this implies that $\mathrm{Adj}_{w'_2} \subseteq \mathrm{Adj}_{w'_1}$. However, as one can easily see from the structure of G^*, it never happens that $\mathrm{Adj}_{w'_2} \subseteq \mathrm{Adj}_{w'_1}$ unless $w'_1 = w'_2$. Thus, we conclude that the map constructed is a DPE of G^* to $\mathbb{R}^3$. With the DPE in hand, by applying Lemmas 46 and 47, the first statement of Theorem 43 is readily proved.

4.2.2.3 #P-hardness for $d \geq 7$ with existence probabilities equal to $\frac{1}{2}$

In this section, we prove the second statement of Theorem 43. When the existence probabilities are restricted to be $\frac{1}{2}$, we are no longer able to apply the tricks used in the previous section, as the reduction from the CSISC problem (Lemma 46) cannot be done under such a restriction. This is the reason for why we have to "loosen" the dimension to 7 in this case.

As we have seen, for a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ with $\pi(a) = \frac{1}{2}$ for all $a \in S$, computing $\mathit{\Gamma}_{\mathcal{S}}$ is totally equivalent to counting independent sets for $G_{\mathcal{S}}$. Therefore, we complete the proof by establishing a more direct reduction from independent-set counting for 3-regular planar bipartite graphs, which constructs directly a DPE of the input graph to $\mathbb{R}^7$. However, it is non-obvious that any 3-regular planar bipartite graph G has dimension at most 7 and how to construct a DPE of G to $\mathbb{R}^7$ in polynomial time. To prove this, we introduce a new technique based on graph coloring. Indeed, we consider a more general case in which the graph G is an arbitrary bipartite graph. The graph coloring to be used is slightly different from the conventional notion, which we call *halfcoloring*. Let $G = (V \cup V', E)$ be a bipartite graph. For any two distinct vertices $u, v \in V$, we define $u \sim v$ if there exists a vertex in V' adjacent to both u and v.

Definition 51. *A k-halfcoloring of G on V is a map $h : V \to \{1, \ldots, k\}$. The halfcoloring h is said to be discrete if $h(u) \neq h(v)$ for any $u, v \in V$ with $u \sim v$, to be semi-discrete if it satisfies the condition that for any distinct $u, v, w \in V$ with $u \sim v$ and $v \sim w$, $h(u), h(v), h(w)$ are not all the same. Symmetrically, we may also*

define a k-halfcoloring on V'.

We may relate halfcoloring to the conventional graph coloring as follows. Define $G' = (V, E')$ with $E' = \{(u, v) : u \sim v \text{ in } G\}$. Clearly, a discrete k-halfcoloring of G on V corresponds to a (conventional) k-coloring of G' satisfying the condition that no two adjacent vertices share the same color, i.e., the subgraph of G' induced by each color form an independent set of G'. Similarly, a semi-discrete k-halfcoloring of G on V corresponds to a k-coloring of G' satisfying that the subgraph of G' induced by each color consists of connected components of sizes at most 2. If h is a k-halfcoloring of G on V, then for each $v' \in V'$ we denote by $\chi_h(v')$ the number of the colors "adjacent" to v' (the color i is said to be adjacent to v' if there is a vertex $v \in V$ adjacent to v' with $h(v) = i$). The following theorem establishes a relation between halfcoloring and graph dimension.

Theorem 52. *Let $G = (V \cup V', E)$ be a bipartite graph.*
***(i)** If there exists a semi-discrete k-halfcoloring $h : V \to \{1, \ldots, k\}$ of G (on V), then $\dim(G) \leq 2k$. Furthermore, with h in hand, one can compute in polynomial time a DPE of G to $\mathbb{R}^{2k}$.*
***(ii)** If, in addition to **(1)**, we have $\chi_h(v') < k$ for all $v' \in V'$, then $\dim(G) \leq 2k-1$. Also, with h in hand, one can compute in polynomial time a DPE of G to $\mathbb{R}^{2k-1}$.*

Proof. Suppose $n = |V \cup V'|$. Let $h : V \to \{1, \ldots, k\}$ be a semi-discrete k-halfcoloring of G (on V). We show $\dim(G) \leq 2k$ by explicitly constructing a DPE $f : V \cup V' \to \mathbb{R}^{2k}$ of G. For $i \in \{1, \ldots, k\}$, we define $V_i = h^{-1}(\{i\}) \subseteq V$ (i.e., V_i consists of the vertices in V colored with color i by h) and define G_i as the subgraph of G with the vertex set $V_i \cup V'$. We first construct k functions $f_1, \ldots, f_k : V \cup V' \to \mathbb{R}^2$, and then obtain the DPE f by identifying $\mathbb{R}^{2k}$ with $(\mathbb{R}^2)^k$ and "combining" the functions $f_1, \ldots, f_k$, i.e., setting

$$f(v) = (f_1(v), \ldots, f_k(v))$$

for all $v \in V \cup V'$. Fixing $p \in \{1, \ldots, k\}$, we describe the construction of f_p. Suppose the graph G_p consists of m connected components. For each $i \in \{1, \ldots, m\}$, let C_i be the set of the vertices in the i-th connected component of G_p. Also, for each $i \in \{1, \ldots, m\}$, let

$$B_i = \{(x, y) \in \mathbb{R}^2 : i - 1 < x < i, m - i < y < m - i + 1\}$$

be an open box in $\mathbb{R}^2$ (see the left part of Figure 4.10). The function f_p to be constructed maps the vertices in C_i to points in B_i as follows. Since h is semi-discrete, we know that $|C_i \cap V| \leq 2$. If $|C_i \cap V| = 0$, then C_i only contains an isolated vertex $v' \in V'$, and we set $f_p(v')$ to be an arbitrary point in B_i. If $|C_i \cap V| = 1$, let v be the only vertex in $C_i \cap V$ and suppose $C_i \cap V' = \{v'_1, \ldots, v'_r\}$. In this case, we set $f_p(v'_1), \ldots, f_p(v'_r)$ to be a sequence of r points in B_i with increasing x-coordinates and decreasing y-coordinates, and $f_p(v)$ to be an arbitrary point in B_i dominated by all of $f_p(v'_1), \ldots, f_p(v'_r)$. See the middle part of Figure 4.10 for an intuitive illustration for this case. If $|C_i \cap V| = 2$, let v_1, v_2 be the two vertices in $C_i \cap V$ and again suppose $C_i \cap V' = \{v'_1, \ldots, v'_r\}$. We may assume that the vertices in $C_i \cap V'$ adjacent to v_1 (resp., v_2) are exactly $v'_1, \ldots, v'_\alpha$ (resp., $v'_\beta, \ldots, v'_r$) for some $\alpha, \beta \in \{1, \ldots, r\}$ with $\alpha \geq \beta$ (if not, one can easily relabel the points to achieve this). Again, we set $f_p(v'_1), \ldots, f_p(v'_r)$ to be a sequence of r points in B_i with increasing x-coordinates and decreasing y-coordinates. Then we set $f_p(v_1)$ to be a point in B_i which is dominated by exactly $f_p(v'_1), \ldots, f_p(v'_\alpha)$, and set $f_p(v_2)$ to be a point in B_i which is dominated by exactly $f_p(v'_\beta), \ldots, f_p(v'_r)$. Note that we can definitely find such two points, since $f_p(v'_1), \ldots, f_p(v'_r)$ have increasing x-coordinates and decreasing y-coordinates. In addition, by carefully determining the locations of $f_p(v_1)$ and $f_p(v_2)$ in B_i, we may further require that $f_p(v_1)$ and $f_p(v_2)$ do not dominate each other. See the right part of Figure 4.10 for an intuitive illustration for this case. After considering all C_i, the function f_p is defined for all vertices in $V_p \cup V'$ (which is the vertex set of G_p). So it suffices to define f_p on $V \backslash V_p$. For each $v \in V \backslash V_p$, we simply set $f_p(v)$ to be an arbitrary point in the box $[-N, -N+1] \times [-N, -N+1]$ for a sufficiently large integer $N > 10n$ (recall that $n = |V \cup V'|$), which completes the construction of f_p. We observe that f_p has the following properties.

(1) For any $v \in V$ and $w \in V_p$, $f_p(v) \not\succ f_p(w)$.

(2) For any $v' \in V'$, $f_p(v')$ is not dominated by any point in the image of f_p.

(3) For any $v \in V_p$ and $v' \in V'$, $f_p(v') \succ f_p(v)$ iff v and v' are adjacent in G.

We do the same thing for all $p \in \{1, \ldots, k\}$ and obtain the functions $f_1, \ldots, f_k$. As mentioned before, we then define $f : V \cup V' \to \mathbb{R}^{2k}$ as $f(v) = (f_1(v), \ldots, f_k(v))$. We now prove that f is a DPE of G. First, for any $v \in V$, we claim that $f(v)$ does not dominate any point in the image of f. Indeed, $f(v) \not\succ f(v')$ for any $v' \in V'$,

since $f_1(v')$ is not dominated by any point in the image of f_1 by property (2) above. Also, $f(v) \not\succ f(w)$ for any $w \in V$, since $f_p(v) \not\succ f_p(w)$ for $p = h(w)$ by property (1) above. Second, for any $v' \in V'$, we have that $f(v')$ is not dominated by any point in the image of f, simply because $f_1(v')$ is not dominated by any point in the image of f_1 by property (2) above. Finally, consider two vertices $v \in V$ and $v' \in V'$. We claim that $f(v') \succ f(v)$ iff v and v' are adjacent in G. If v and v' are adjacent, then $f_i(v') \succ f_i(v)$ for all $i \in \{1, \ldots, k\}$ by property (3) above, and hence $f(v') \succ f(v)$. If v and v' are not adjacent, then $f_p(v') \not\succ f_p(v)$ for $p = h(v)$ by property (3) above, and hence $f(v') \not\succ f(v)$. In sum, we have $f(v') \succ f(v)$ iff $v \in V$, $v' \in V'$, v and v' are adjacent in G. Therefore, f is a DPE of G to $\mathbb{R}^{2k}$. Clearly, f can be constructed in polynomial time if the k-halfcoloring h is provided, which completes the proof of the first part of the theorem.

Next, we prove the second part of the theorem. Again, let $h : V \to \{1, \ldots, k\}$ be a semi-discrete k-halfcoloring of G (on V). Suppose $\chi_h(v') < k$ for all $v' \in V'$. If $k = 1$, then $\chi_h(v') = 0$ for all $v' \in V'$, which implies that G has no edges and thus the statement is trivial (any constant map $f : V \cup V' \to \mathbb{R}$ is a DPE of G). So assume $k \geq 2$. We show $\dim(G) \leq 2k - 1$ by explicitly constructing a DPE $f : V \cup V' \to \mathbb{R}^{2k-1}$ of G. In the same way as before, we define the functions $f_1, \ldots, f_k : V \cup V' \to \mathbb{R}^2$. But we need a different way to define f. To this end, we first construct $k - 1$ functions $f'_1, \ldots, f'_{k-1} : V \cup V' \to \mathbb{R}^2$ based on $f_1, \ldots, f_k$ as follows. Fixing $p \in \{1, \ldots, k - 1\}$, we describe the construction of f'_p. For all $v \in V \backslash V_k$, we set $f'_p(v) = f_p(v)$. For all $v \in V_k$, we set $f'_p(v) = f_k(v) - (n, n)$, that is, if $f_k(v) = (x, y) \in \mathbb{R}^2$ then $f'_p(v) = (x - n, y - n)$. Now consider the vertices in V'. If a vertex $v' \in V'$ is "adjacent" to the color p (recall that v' is said to be "adjacent" to the color p if there exists $v \in V$ adjacent to v' with $h(v) = p$),

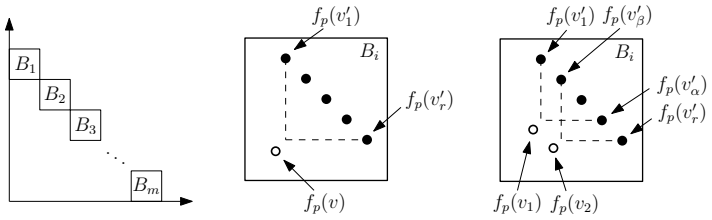


Figure 4.10: A local structure of f_p in the box B_i.

then we set $f'_p(v') = f_p(v')$, otherwise $f'_p(v') = f_k(v') - (n, n)$. By doing this for all $p \in \{1, \ldots, k-1\}$, we complete constructing $f'_1, \ldots, f'_{k-1}$. However, if we "combine" $f'_1, \ldots, f'_{k-1}$, we only obtain a map $V \cup V' \to \mathbb{R}^{2k-2}$ which is not guaranteed to be a DPE. So the last ingredient needed for defining f is a function $\rho : V \cup V' \to \mathbb{R}$. The definition of ρ is quite simple. We set $\rho(v) = 1$ for all $v \in V \backslash V_k$, and $\rho(v) = 3$ for all $v \in V_k$. For $v' \in V'$, if v' is "adjacent" to the color k or $\chi_h(v') = 0$, then we set $\rho(v') = 4$, otherwise $\rho(v') = 2$. Finally, $f : V \cup V' \to \mathbb{R}^{2k-1}$ is defined by identifying $\mathbb{R}^{2k-1}$ with $(\mathbb{R}^2)^{k-1} \times \mathbb{R}$ and "combining" the functions $f'_1, \ldots, f'_{k-1}, \rho$, i.e., setting

$$f(v) = (f'_1(v), \ldots, f'_{k-1}(v), \rho(v))$$

for all $v \in V \cup V'$. We need to verify that f is truly a DPE of G to $\mathbb{R}^{2k-1}$.

First, we show that for any $v \in V$, $f(v)$ does not dominate any point in the image of f. Let $v \in V$ be a vertex. We consider two cases, $v \in V \backslash V_k$ and $v \in V_k$. In the case of $v \in V \backslash V_k$, we first notice that $f(v) \not\succ f(w)$ for any $w \in V_k \cup V'$, simply because $\rho(v) < \rho(w)$. To see this $f(v) \not\succ f(w)$ for any $w \in V \backslash V_k$, set $p = h(w) \neq k$. Then $f'_p(v) = f_p(v)$ does not dominate $f'_p(w) = f_p(w)$ by property (1) above, and hence $f(v) \not\succ f(w)$. In the case of $v \in V_k$, we first claim that $f(v) \not\succ f(w)$ for any $w \in V$. If $w \notin V_k$, then by setting $p = h(w) \neq k$ we have $f'_p(v) = f_k(v) - (n, n)$ does not dominate $f'_p(w) = f_p(w)$, which implies $f(v) \not\succ f(w)$. If $w \in V_k$, then $f'_1(v) = f_k(v) - (n, n)$ does not dominate $f'_1(w) = f_k(w) - (n, n)$ since $f_k(v) \not\succ f_k(w)$ by property (1) above, which also implies $f(v) \not\succ f(w)$. It suffices to show that $f(v) \not\succ f(v')$ for any $v' \in V'$. Indeed, we have either $f'_1(v') = f_1(v')$ or $f'_1(v') = f_k(v') - (n, n)$. In each case, $f'_1(v) = f_k(v) - (n, n)$ does not dominate $f'_1(v')$ (the former case is obvious and the latter case follows from property (2) above). Thus $f(v) \not\succ f(v')$.

Second, we show that for any $v' \in V'$, $f(v')$ is not dominated by any point in the image of f. Let $v' \in V'$ be a vertex. By the argument above, it suffices to verify that $f(w') \not\succ f(v')$ for any $w' \in V'$. If v' is "adjacent" to some color $p \in \{1, \ldots, k-1\}$, then we are done because $f'_p(v') = f_p(v')$ is not dominated by $f'_p(w')$ for any $w' \in V'$. Suppose v' is not "adjacent" to any color in $\{1, \ldots, k-1\}$. In this case, we must have $\rho(v') = 4$ and $f'_i(v') = f_k(v') - (n, n)$ for all $i \in \{1, \ldots, k-1\}$. We first notice that $f(w') \not\succ f(v')$ for any $w' \in V'$ such that $\chi_h(w') > 0$ and w' is not "adjacent" to the color k, simply because $\rho(w') = 2 < \rho(v')$. Then we consider the case that

$w' \in V'$ is "adjacent" to the color k or $\chi_h(w') = 0$. By the assumption $\chi_h(w') < k$, we know that w' cannot be "adjacent" to all the k colors. In other words, if w' is "adjacent" to the color k or $\chi_h(w') = 0$, w' must miss some color in $\{1, \ldots, k-1\}$. Without loss of generality, we may assume w' is not "adjacent" to the color 1. Thus, $f'_1(v') = f_k(v') - (n, n)$ is not dominated by $f'_1(w') = f_k(w') - (n, n)$ by property (2) above, and hence $f(w') \not\succ f(v')$.

Finally, we show that for any $v \in V$ and $v' \in V'$, $f(v') \succ f(v)$ iff v and v' are adjacent in G. Let $v \in V$ and $v' \in V'$ be two vertices. If v and v' are adjacent in G, one can easily verify (by checking various cases) that $\rho(v') > \rho(v)$ and $f'_i(v')$ dominates $f'_i(v)$ for all $i \in \{1, \ldots, k-1\}$, which implies $f(v') \succ f(v)$. Now suppose v and v' are not adjacent in G. We consider two cases, $v \in V \backslash V_k$ and $v \in V_k$. In the case of $v \in V \backslash V_k$, set $p = h(v) \neq k$. Then $f'_p(v) = f_p(v)$. Besides, we have either $f'_p(v') = f_p(v')$ or $f'_p(v') = f_k(v') - (n, n)$. For the former, $f'_p(v') \not\succ f'_p(v)$ follows from property (3) above, while for the latter $f'_p(v') \not\succ f'_p(v)$ follows obviously. Thus, $f(v') \not\succ f(v)$. In the case of $v \in V_k$, we have $f'_i(v) = f_k(v) - (n, n)$ for all $i \in \{1, \ldots, k\}$ and $\rho(v) = 3$. If v' is not "adjacent" to the color k and $\chi_h(v') > 0$, then $\rho(v') = 2 < \rho(v)$ and hence $f(v') \not\succ f(v)$. If v' is "adjacent" to the color k or $\chi_h(v') = 0$, then as argued before v' must miss some color in $\{1, \ldots, k-1\}$. Without loss of generality, we may assume w' is not "adjacent" to the color 1. Thus, $f'_1(v') = f_k(v') - (n, n)$ does not dominate $f'_1(v) = f_k(v) - (n, n)$ by property (3) above, which implies $f(v') \not\succ f(v)$.

In sum, two vertices in G share a common edge iff their images under f form a dominance. Therefore, f is a DPE of G to $\mathbb{R}^{2k-1}$. It is clear that the construction of f can be done in polynomial time if the k-halfcoloring h is provided. □

We then apply the halfcoloring technique to show that $\dim(G) \leq 7$ for any 3-regular planar bipartite graph G, which will give us a proof for the second statement of Theorem 43. To achieve this, the only missing piece is the following observation.

Lemma 53. *Every 3-regular planar bipartite graph has a discrete 4-halfcoloring, which can be computed in polynomial time.*

Proof. Let $G = (V \cup V', E)$ be a 3-regular planar bipartite graph. As before, we define the graph $G' = (V, E')$ by setting $E' = \{(a, b) : a \sim b \text{ in } G\}$. Then a discrete

k-halfcoloring of G on V corresponds to a (conventional) k-coloring of G' satisfying that no two adjacent vertices share the same color. We first show that G' is planar. Fix a planar drawing φ of G. Let $v' \in V'$ be a vertex. Since G is 3-regular, v' must be adjacent to three vertices $v_1, v_2, v_3 \in V$. We now delete v' as well as its three adjacent edges from G and add three new edges $(v_1, v_2), (v_2, v_3), (v_3, v_1)$ to G. We claim that the resulting graph is still planar. Indeed, in the drawing φ, after we remove $\varphi(v')$ and its adjacent edges, $\varphi(v_1), \varphi(v_2), \varphi(v_3)$ will share a common face, which is the one previously containing $\varphi(v)$. So we can draw the edges $(v_1, v_2), (v_2, v_3), (v_3, v_1)$ inside this face along with the image of the deleted edges (see Figure 4.11). In this way, we keep deleting the vertices in V' (as well as the adjacent edges) and adding new edges. In this process, the planarity of the graph is always maintained. When all the vertices in V' are deleted, the resulting graph, which is still planar, is nothing but G', as two vertices $u, v \in V$ are connected (in the resulting graph) iff $u \sim v$ in G. By applying the well-known Four Color Theorem, we know that G' is 4-colorable. Furthermore, to find a 4-coloring for G' can be done in quadratic time using the approach in [37]. As a result, a discrete 4-halfcoloring of G can be computed in polynomial time, completing the proof. □

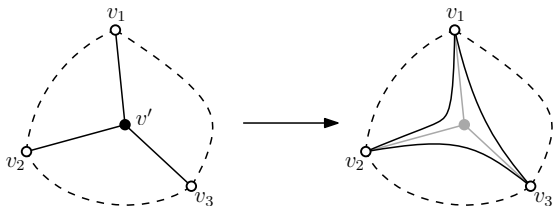


Figure 4.11: Deleting a vertex and adding three new edges.

Now it is quite straightforward to prove the second statement of Theorem 43. Let G be a 3-regular planar bipartite graph. By combining Theorem 52 and Lemma 53, we can compute a DPE of G to $\mathbb{R}^7$ in polynomial time. By taking the images of the vertices of G under the DPE, we obtain a set S of points in $\mathbb{R}^7$. Using the point set S, we further construct a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ by choosing an injection $\mathrm{cl} : S \to \mathbb{N}$ and defining $\pi(a) = \frac{1}{2}$ for any $a \in S$. It is clear that $G_{\mathcal{S}} \cong G$ and thus $Ind(G) = 2^{|S|} \Gamma_{\mathcal{S}}$. Then by applying Lemma 47, we can compute another colored stochastic dataset $\mathcal{S}' = (S', \mathrm{cl}', \pi')$ such that $\Gamma_{\mathcal{S}'} = \Gamma_{\mathcal{S}}$ and $\pi'(a) = \frac{1}{2}$ for any

$a \in S'$, and more importantly, $\langle \mathcal{S}' \rangle$ is an instance of the CSD problem with respect to $\mathcal{P}$. With this reduction, the second statement of Theorem 43 is proved.

4.2.3 A simple FPRAS

In this section, we describe a simple FPRAS (i.e., fully polynomial-time randomized approximation scheme) for approximating $\Lambda_{\mathcal{S}}$ in any dimension. Recall that a FPRAS is a randomized algorithm which takes the input of the problem with an additional parameter $\varepsilon > 0$, and computes an ε-approximation of the answer in polynomial (in both the size of the problem and $1/\varepsilon$) time with high probability (say at least 2/3).

A natural idea to design a FPRAS for approximating $\Lambda_{\mathcal{S}}$ is to randomly generate a large number of realizations of $\mathcal{S}$, and estimate $\Lambda_{\mathcal{S}}$ using the proportion of the number of the realizations containing inter-color dominances to the total number of the realizations. However, since we are only allowed to generate a polynomial number of realizations, this method does not guarantee to produce an ε-approximation of $\Lambda_{\mathcal{S}}$ with high probability. For instance, if $\Lambda_{\mathcal{S}} = 2^{-n}$, then the estimation of $\Lambda_{\mathcal{S}}$ obtained by generating polynomial number of realizations would be 0 with probability almost 1 (as one can easily verify using union bound). Interestingly, by slightly making some changes to this simple method, we can truly obtain a FPRAS for computing $\Lambda_{\mathcal{S}}$.

Our FPRAS works as follows. Suppose the points $a_1, \ldots, a_n$ are already sorted by their existence probabilities from large to small, i.e., $\pi(a_1) \geq \cdots \geq \pi(a_n)$. Instead of estimating $\Lambda_{\mathcal{S}}$ directly, what we do is to estimate a set of conditional probabilities and use them to compute an estimation of $\Lambda_{\mathcal{S}}$. For any $i, j \in \{1, \ldots, n\}$ with $i < j$, we define $E_{i,j}$ as the event that a realization R of $\mathcal{S}$ includes a_i, a_j and any other points in R have indices smaller than i. Then we immediately have

$$\Lambda_{\mathcal{S}} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Pr[E_{i,j}] \cdot \mathit{Cond}_{i,j}, \tag{4.2}$$

where $\mathit{Cond}_{i,j}$ is the conditional probability that a realization of $\mathcal{S}$ contains inter-color dominances under the condition that $E_{i,j}$ happens. The probabilities $\Pr[E_{i,j}]$ can be straightforwardly computed. But we are not able to exactly compute $\mathit{Cond}_{i,j}$ in polynomial time, so we try to estimate them by randomly generating realizations.

For $p \in \{0, 1, \ldots, n\}$, set $S_p = \{a_1, \ldots, a_p\}$, and we use $\mathcal{S}_p$ to denote the sub-dataset of $\mathcal{S}$ with point set $S_p \subseteq S$. We randomly generate $N = 10n^5/\varepsilon^2$ realizations of $\mathcal{S}_p$ for each $p \in \{0, 1, \ldots, n\}$. Let $R_{p,q}$ be the q-th realization of $\mathcal{S}_p$. We compute an estimation $Est_{i,j}$ for each $Cond_{i,j}$ as

$$Est_{i,j} = \sum_{k=1}^{N} \frac{\sigma(R_{i-1,k} \cup \{a_i, a_j\})}{N},$$

where $\sigma(R) = 1$ if R contains inter-color dominances and $\sigma(R) = 0$ otherwise. Then we can apply Equation 4.2 to compute an estimation Λ of $\Lambda_{\mathcal{S}}$, simply by replacing each $Cond_{i,j}$ with its estimation $Est_{i,j}$. It is quite surprising that Λ is, with high probability, an ε-approximation of $\Lambda_{\mathcal{S}}$ (note that each $Est_{i,j}$ is not necessarily an ε-approximation of $Cond_{i,j}$ with high probability). The following theorem completes the discussion.

Theorem 54. *We have $(1-\varepsilon)\Lambda_{\mathcal{S}} < \Lambda < (1+\varepsilon)\Lambda_{\mathcal{S}}$ with probability at least $2/3$.*

Proof. We show that for any $i, j \in \{1, \ldots, n\}$ with $i < j$,

$$\Pr[E_{i,j}] \cdot |Est_{i,j} - Cond_{i,j}| < \frac{\varepsilon}{n^2}\Lambda_{\mathcal{S}} \tag{4.3}$$

with probability $1 - O(e^{-n})$. As long as this is true, by using union bound, we can immediately conclude that $|\Lambda - \Lambda_{\mathcal{S}}| < \varepsilon\Lambda_{\mathcal{S}}$ with probability at least $2/3$, which completes the proof. Consider a realization R of $\mathcal{S}_{i-1}$. Clearly, the probability that $R \cup \{a_i, a_j\}$ contains inter-color dominances is nothing but $Cond_{i,j}$. Therefore, by Hoeffding's inequality and the definition of $Est_{i,j}$, we have that

$$\Pr\left[|Est_{i,j} - Cond_{i,j}| \geq \frac{\varepsilon}{n^2}\right] \leq 2e^{-2N\varepsilon^2/n^4} = 2e^{-2n}.$$

If $\Pr[E_{i,j}] \leq \Lambda_{\mathcal{S}}$, we are done because the above already implies that Inequality 4.3 holds with probability $1-O(e^{-n})$. So assume $\Pr[E_{i,j}] > \Lambda_{\mathcal{S}}$. Note that $\pi(a_i)\cdot\pi(a_j) \geq \Pr[E_{i,j}]$, which implies $\pi(a_i) \cdot \pi(a_j) > \Lambda_{\mathcal{S}}$. We claim that $Cond_{i,j} = 0$. It suffices to show that for any realization R of $\mathcal{S}_{i-1}$, $R \cup \{a_i, a_j\}$ contains no inter-color dominances. Let $a_p, a_q \in R \cup \{a_i, a_j\}$ be two distinct points. Assume $\mathrm{cl}(a_p) \neq \mathrm{cl}(a_q)$ and $a_p \succ a_q$. Then we must have $\Lambda_{\mathcal{S}} \geq \pi(a_p) \cdot \pi(a_q)$ because a realization of $\mathcal{S}$ does contain inter-color dominances if it includes both a_p and a_q. However, recall that $\pi(a_1) \geq \cdots \geq \pi(a_n)$. Thus, $\pi(a_p) \cdot \pi(a_q) \geq \pi(a_i) \cdot \pi(a_j) > \Lambda_{\mathcal{S}}$, which gives us a

contradiction. Since $Cond_{i,j} = 0$, $Est_{i,j}$ is for sure 0. It follows that Inequality 4.3 holds with probability 1 in this case. As a result, $(1-\varepsilon)\Lambda_{\mathcal{S}} < \Lambda < (1+\varepsilon)\Lambda_{\mathcal{S}}$ with probability at least 2/3. □

4.3 The free-basis colored stochastic dominance problem

Define $\Lambda^*_{\mathcal{S}}$ as the probability that a realization of $\mathcal{S}$ contains inter-color dominances with respect to any orthogonal basis of $\mathbb{R}^d$. Set $\Gamma^*_{\mathcal{S}} = 1 - \Lambda^*_{\mathcal{S}}$, which is the probability that a realization of $\mathcal{S}$ contains no inter-color dominances with respect to some orthogonal basis of $\mathbb{R}^d$. The goal of the FBCSD problem is to compute $\Lambda^*_{\mathcal{S}}$ (or $\Gamma^*_{\mathcal{S}}$).

4.3.1 Reduction from the CSD problem

In this section, we show that the (standard) CSD problem in $\mathbb{R}^d$ is polynomial-time reducible to the FBCSD problem in the same dimension, which implies the latter is #P-hard for $d \geq 3$. Given a colored stochastic dataset $\mathcal{S} = (S, \mathrm{cl}, \pi)$ in $\mathbb{R}^d$ as an instance of the CSD problem, our reduction tries to construct another colored stochastic dataset $\mathcal{S}' = (S', \mathrm{cl}', \pi')$ in $\mathbb{R}^d$ such that $\Lambda^*_{\mathcal{S}'} = \Lambda_{\mathcal{S}}$. The intuition for our reduction is the following. First, consider the given colored stochastic dataset $\mathcal{S}$. Clearly, we have $\Lambda^*_{\mathcal{S}} \leq \Lambda_{\mathcal{S}}$, as every realization of $\mathcal{S}$ counted in $\Lambda^*_{\mathcal{S}}$ is also counted in $\Lambda_{\mathcal{S}}$. The reason for why $\Lambda^*_{\mathcal{S}}$ may be smaller than $\Lambda_{\mathcal{S}}$ is that perhaps some realization contains inter-color dominances with respect to the standard basis E of $\mathbb{R}^d$ but does not contain inter-color dominances with respect to some other basis. To handle this, our basic idea is to add a set Ψ of (colored) auxiliary points with existence probabilities 1 to S, that is, we want $S' = S \cup \Psi$ with $\pi'(b) = 1$ for all $b \in \Psi$ (and $\pi'(a) = \pi(a)$, $\mathrm{cl}'(a) = \mathrm{cl}(a)$ for all $a \in S$). The goal of adding these auxiliary points is to guarantee that a subset $A \subseteq S$ contains inter-color dominances with respect to the standard basis E iff $A \cup \Psi \subseteq S'$ contains inter-color dominances with respect to any orthogonal basis. Note that as long as Ψ has this property, it obviously holds that $\Lambda^*_{\mathcal{S}'} = \Lambda_{\mathcal{S}}$. Therefore, the critical part of our reduction is to construct such a set Ψ with the desired property. We achieve this through several steps.

First of all, we need to make the point set S "regular". Formally, we say a

(finite) point set $X \subset \mathbb{R}^d$ is *regular* if $X \subset \{1, 2, \ldots, |X|\}^d$ and any two distinct points $x, x' \in X$ have distinct coordinates in all dimensions. It is easy to see that one can always "regularize" a point set without changing the dominance relation (with respect to E) among the points.

Lemma 55. *Given a set $S = \{a_1, \ldots, a_n\} \subset \mathbb{R}^d$ of distinct points, one can construct in $O(n \log n)$ time a regular set $S_{new} = \{\hat{a}_1, \ldots, \hat{a}_n\} \subset \mathbb{R}^d$ such that $\hat{a}_i \succ_E \hat{a}_j$ iff $a_i \succ_E a_j$.*

Proof. Fixing $p \in \{1, \ldots, d\}$, we determine the p-th coordinates of $\hat{a}_1, \ldots, \hat{a}_n$ as follows. For all $i \in \{1, \ldots, n\}$, define a triple $\phi_i = (\gamma_i, \sigma_i, i)$ where γ_i is the p-th coordinate of a_i and σ_i is the sum of the d coordinates of a_i. Then we sort all ϕ_i in lexicographic order from small to large, and suppose $\phi_{i_1}, \ldots, \phi_{i_n}$ is the resulting sorted sequence. We have $\phi_{i_1} < \cdots < \phi_{i_n}$ under lexicographic order, since there exist no ties. Now we simply set the p-th coordinates of $\hat{a}_{i_1}, \ldots, \hat{a}_{i_n}$ to be $1, \ldots, n$ respectively. In this way, we obtain the new set $S_{new} = \{\hat{a}_1, \ldots, \hat{a}_n\} \subset \mathbb{R}^d$ in $O(n \log n)$ time (note that d is assumed to be constant). It is clear that S_{new} is regular. We verify that S_{new} satisfies the desired property. Assume $a_i \succ_E a_j$. Then in each dimension, the coordinate of a_i is greater than or equal to the coordinate of a_j. In addition, the sum of the d coordinates of a_i is greater than that of a_j. Therefore, in all dimensions, the coordinates of $\hat{a}_i$ are greater than the coordinates of $\hat{a}_j$, i.e., $\hat{a}_i \succ_E \hat{a}_j$. Assume $a_i \not\succ_E a_j$. Then there exists $p \in \{1, \ldots, d\}$ such that the p-th coordinate of a_i is smaller than the p-th coordinate of a_j. By definition, the p-th coordinate of $\hat{a}_i$ is also smaller than the p-th coordinate of $\hat{a}_j$. Therefore, $\hat{a}_i \not\succ_E \hat{a}_j$. □

Now we may assume S is regular. To construct Ψ, we need to introduce some new notions.

Definition 56. *Let $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ be an orthogonal basis of $\mathbb{R}^d$. We define the cone C_B of B as*

$$C_B = \left\{ \sum_{i=1}^{d} \beta_i \mathbf{b}_i : \beta_1, \ldots, \beta_d \geq 0 \right\} \cup \left\{ \sum_{i=1}^{d} \beta_i \mathbf{b}_i : \beta_1, \ldots, \beta_d \leq 0 \right\} \subset \mathbb{R}^d.$$

Also, we define the projective cone $PC_B \subset \mathbb{P}^{d-1}$ as the image of $C_B \backslash \{\mathbf{0}\}$ in $\mathbb{P}^{d-1}$ under the standard quotient map $\mathbb{R}^d \backslash \{\mathbf{0}\} \to \mathbb{P}^{d-1}$.

Intuitively, the cone C_B consists of the points whose coordinates are all positive or all negative under the basis B, and the projective cone PC_B consists of all lines through the origin that lie in C_B. For a point $x \in \mathbb{R}^d$ with $x \neq \mathbf{0}$, we denote by $\overline{x}$ its image in $\mathbb{P}^{d-1}$ under the quotient map $\mathbb{R}^d \backslash \{\mathbf{0}\} \to \mathbb{P}^{d-1}$. The notion of (projective) cone defined above gives us another way to view dominance relations with respect to an orthogonal basis. Consider two distinct points $p, q \in \mathbb{R}^d$, and an orthogonal basis B of $\mathbb{R}^d$. It is easy to see that p, q form a dominance with respect to B (i.e., $p \succ_B q$ or $q \succ_B p$) iff $p - q \in C_B$, or equivalently, $\overline{p-q} \in PC_B$. Another notion we need is a metric on any projective space $\mathbb{P}^k$.

Definition 57. *For two points $l, l' \in \mathbb{P}^k$, we define $\mathrm{ang}(l, l') \in [0, \frac{\pi}{2}]$ to be the angle between l and l' as lines in $\mathbb{R}^{k+1}$ through the origin (there are two supplementary angles, take the smaller one which is in $[0, \frac{\pi}{2}]$). It is easy to see that $\mathrm{ang}(\cdot, \cdot)$ defines a metric on $\mathbb{P}^k$.*

The following lemmas establish some geometric properties of the projective cone and the ang-metric, which will be helpful for constructing Ψ.

Lemma 58. *Let B be an orthogonal basis of $\mathbb{R}^d$, and l be a point in $\mathbb{P}^{d-1}$. If $l \notin PC_B$, then there exists $x \in PC_B$ perpendicular to l, i.e., $\mathrm{ang}(l, x) = \frac{\pi}{2}$.*

Proof. Without loss of generality, we may assume $B = E$. Let $[r_1 : \cdots : r_d]$ be the homogeneous coordinates of l. Since $l \notin PC_B$, we may find r_p and r_q such that $r_p > 0$ and $r_q < 0$. Now we define $r'_1, \ldots, r'_d \in \mathbb{R}$ by setting $r'_p = -r_q$, $r'_q = r_p$, and $r'_i = 0$ for any $i \notin \{p, q\}$. Consider the point $x = [r'_1 : \cdots : r'_d] \in \mathbb{P}^{d-1}$. Note that r'_p and r'_q are nonzero so that x is well-defined. Since $r'_1, \ldots, r'_d$ are nonnegative, we have $x \in PC_B$. Furthermore, we know that $\mathrm{ang}(l, x) = \frac{\pi}{2}$, because $\sum_{i=1}^d r_i r'_i = 0$. □

Lemma 59. *For any orthogonal basis B of $\mathbb{R}^d$, any point $x \in PC_B$, and any real number $\varepsilon \in (0, \frac{\pi}{2}]$, there exists $y \in PC_B$ with $\mathrm{ang}(x, y) < \varepsilon$ such that the $\frac{\varepsilon}{3\sqrt{d}}$-ball at y, i.e., the set $\{z \in \mathbb{P}^{d-1} : \mathrm{ang}(z, y) \leq \frac{\varepsilon}{3\sqrt{d}}\}$, is contained in PC_B.*

Proof. Without loss of generality, we may assume $B = E$. Let $[r_1 : \cdots : r_d]$ be the homogeneous coordinates of x such that $\sum_{i=1}^d r_i^2 = 1$. Since $x \in PC_B$, the coordinates can be chosen such that $r_1, \ldots, r_d$ are nonnegative. Consider the point

$y = [r'_1 : \cdots : r'_d] \in \mathbb{P}^{d-1}$ where $r'_i = r_i + \frac{\varepsilon}{\sqrt{d}}$. It is clear that y is well-defined and in PC_B. Set $\theta = \mathrm{ang}(x, y)$. To see $\theta < \varepsilon$, we note that

$$\sin^2\theta = 1 - \cos^2\theta = 1 - \frac{(\sum_{i=1}^d r_i r'_i)^2}{\sum_{i=1}^d (r'_i)^2} = \frac{(d-\gamma^2)\varepsilon^2}{d + 2\varepsilon\sqrt{d}\gamma + d\varepsilon^2},$$

where $\gamma = \sum_{i=1}^d r_i \geq 1$. Therefore, $\sin^2\theta < \varepsilon^2/(1+\varepsilon^2)$ and $\sin\theta < \varepsilon/\sqrt{1+\varepsilon^2}$, which implies $\theta < \varepsilon$. It suffices to show that the $\frac{\varepsilon}{3\sqrt{d}}$-ball at y is contained in PC_B. Equivalently, we want that $\mathrm{ang}(z, y) > \frac{\varepsilon}{3\sqrt{d}}$ for any $z \in \mathbb{P}^d \backslash PC_B$. Let $z = [s_1 : \cdots : s_d]$ be a point in $\mathbb{P}^d \backslash PC_B$ and assume $\sum_{i=1}^d s_i^2 = 1$. We have that

$$\cos^2(\mathrm{ang}(z, y)) = \frac{\left(\sum_{i=1}^d r'_i s_i\right)^2}{\sum_{i=1}^d (r'_i)^2}.$$

Because $z \in \mathbb{P}^d \backslash PC_B$, there must exist some p, q such that $s_p > 0$ and $s_q < 0$. Since $r'_1, \ldots, r'_d > 0$ and $\sum_{i=1}^d s_i^2 = 1$, we have that

$$\frac{\left(\sum_{i=1}^d r'_i s_i\right)^2}{\sum_{i=1}^d (r'_i)^2} \leq \frac{\left(\sum_{i=1}^d r'_i |s_i|\right)^2 - \eta^2}{\sum_{i=1}^d (r'_i)^2} \leq \frac{\sum_{i=1}^d (r'_i)^2 - \eta^2}{\sum_{i=1}^d (r'_i)^2},$$

where $\eta = \min\{|r'_p|, |r'_q|\}$. It follows that

$$\sin^2(\mathrm{ang}(z, y)) \geq \frac{\eta^2}{\sum_{i=1}^d (r'_i)^2} \geq \frac{\varepsilon^2/d}{(1+\varepsilon)^2} > \frac{\varepsilon^2}{9d}.$$

Therefore, $\mathrm{ang}(z, y) \geq \sin(\mathrm{ang}(z, y)) > \frac{\varepsilon}{3\sqrt{d}}$. □

Lemma 60. *Let l be a point in $\mathbb{P}^{d-1}$ and $\varepsilon \geq \xi > 0$ be two real numbers. Then one can compute $m = O(\varepsilon/\xi^{d-1})$ points $l_1, \ldots, l_m \in \mathbb{P}^{d-1}$ in $O(m)$ time such that (1) $\mathrm{ang}(l, l_i) > \frac{\pi}{2} - \varepsilon$ for all $i \in \{1, \ldots, m\}$ and (2) for any $y \in \mathbb{P}^{d-1}$ with $\mathrm{ang}(l, y) > \frac{\pi}{2} - \varepsilon$, there exists some l_i satisfying $\mathrm{ang}(l_i, y) < \xi$.*

Proof. By taking $\varepsilon > \frac{\pi}{2}$, the statement in the theorem implies that for any $\xi > 0$, one can compute $m = O(1/\xi^{d-1})$ points $l_1, \ldots, l_m \in \mathbb{P}^{d-1}$ in $O(m)$ time such that $\min_i \mathrm{ang}(l_i, y) < \xi$ for any $y \in \mathbb{P}^{d-1}$. With this observation, we complete the proof by applying induction on the dimension. In $\mathbb{P}^1$, the statement is quite obvious. Without loss of generality, we may assume $l = [0 : 1]$. Set $\gamma = \lfloor \varepsilon/\xi \rfloor$ and $m = 2\gamma + 1$. Then one can simply take the m points $[\cos(i\xi) : \sin(i\xi)]$ for all $i \in \{-\gamma, \ldots, 0, \ldots, \gamma\}$

as $l_1, \ldots, l_m$. The two desired properties of $l_1, \ldots, l_m$ can be readily verified. Now suppose the theorem holds in $\mathbb{P}^{k-1}$, and we consider the case in $\mathbb{P}^k$. Similarly, we may assume $l = [0 : \cdots : 0 : 1] \in \mathbb{P}^k$. As argued at the beginning, our induction hypothesis implies that we can compute $m' = O(1/\xi^{k-1})$ points $l'_1, \ldots, l'_{m'} \in \mathbb{P}^{k-1}$ in $O(m')$ time such that $\min_i \operatorname{ang}(l'_i, y) < \xi/2$ for any $y \in \mathbb{P}^{k-1}$. We then use these m' points to achieve our construction in $\mathbb{P}^k$ as follows. For any real number $\alpha \in [0, 1)$, we define the inclusion map $f_\alpha : \mathbb{P}^{k-1} \to \mathbb{P}^k$ as

$$f_\alpha : [r_1 : \cdots : r_k] \mapsto \left[r_1 : \cdots : r_k : \sqrt{\frac{t}{1-\alpha^2}} \right],$$

where $t = \sum_{i=1}^{k} r_i^2$ (note that f_α is well-defined). Set $\gamma = \lfloor 2\varepsilon/\xi \rfloor$ and $m = (2\gamma + 1)m' = O(\varepsilon/\xi^k)$. Also, set $\alpha_i = \sin(i\xi/2)$ for $i \in \{-\gamma, \ldots, 0, \ldots, \gamma\}$. Then we take the m points $f_{\alpha_i}(l'_j)$ for all $i \in \{-\gamma, \ldots, 0, \ldots, \gamma\}$ and all $j \in \{1, \ldots, m'\}$ as $l_1, \ldots, l_m$. It suffices to show that $l_1, \ldots, l_m$ satisfy the two desired conditions. Clearly, for any $i = \{-\gamma, \ldots, 0, \ldots, \gamma\}$ and $j = \{1, \ldots, m'\}$, we have that $\operatorname{ang}(l, f_{\alpha_i}(l'_j)) = \frac{\pi}{2} - i\xi/2 > \frac{\pi}{2} - \varepsilon$. To verify the condition (2), let $y = [r_1 : \cdots : r_{k+1}]$ be a point in $\mathbb{P}^k$ where $\sum_{i=1}^{k+1} r_i^2 = 1$. Suppose $\operatorname{ang}(l, y) > \frac{\pi}{2} - \varepsilon$, so $|r_{k+1}| < \sin\varepsilon$. If $r_{k+1} \geq 0$, we define p as the largest integer in $\{0, \ldots, \gamma\}$ such that $\sin(p\xi/2) \leq r_{k+1}$, otherwise define p as the smallest integer in $\{-\gamma, \ldots, 0\}$ such that $\sin(p\xi/2) \geq r_{k+1}$. Set $y' = [r_1 : \cdots : r_k] \in \mathbb{P}^{k-1}$. Then by assumption, there exists some $q \in \{1, \ldots, m'\}$ such that $\operatorname{ang}(l'_q, y') < \xi/2$. We claim that $\operatorname{ang}(f_{\alpha_p}(l'_q), y) < \xi$. Indeed, we consider the point $f_{\alpha_p}(y') \in \mathbb{P}^k$. We have $\operatorname{ang}(f_{\alpha_p}(l'_q), f_{\alpha_p}(y')) \leq \operatorname{ang}(l'_q, y') < \xi/2$. Also, we have $\operatorname{ang}(f_{\alpha_p}(y'), y) = |\arcsin(r_{k+1}) - p\xi/2| < \xi/2$. Therefore, $\operatorname{ang}(f_{\alpha_p}(l'_q), y) < \xi$, which implies that the points $l_1, \ldots, l_m$ satisfy the condition (2). The induction argument then completes the proof. □

With the above lemmas in hand, we now describe the construction of Ψ. We look at all pairs (a, a') of points in S such that $\operatorname{cl}(a) \neq \operatorname{cl}(a')$ and $a \succ_E a'$. For each such pair (a, a'), we do the following. Set $l = \overline{a - a'} \in \mathbb{P}^{d-1}$, $\varepsilon = \arcsin(\frac{1}{\sqrt{dn}})$, and $\xi = \frac{\varepsilon}{3\sqrt{d}}$. By applying Lemma 60 with l, ε, ξ, we compute $m = O(\varepsilon/\xi^{d-1}) = O(n^{d-2})$ points $l_1, \ldots, l_m \in \mathbb{P}^{d-1}$ satisfying the conditions (1) and (2) in the lemma. In addition, we observe the following.

- $l_i \notin PC_E$ for all $i \in \{1, \ldots, m\}$.
- For any orthogonal basis B of $\mathbb{R}^d$, if $l \notin PC_B$, then there exists some $l_i \in PC_B$.

To see the first observation, recall that S is already regular. Since $a \succ_E a'$ and S is regular, l can be represented by homogeneous coordinates $[\alpha_1 : \cdots : \alpha_d]$ with $\alpha_1, \ldots, \alpha_d \in \{1, \ldots, n-1\}$. Based on this, one can easily verify that $\text{ang}(l, l') < \arccos(\frac{1}{\sqrt{d}n})$ for any $l' \in PC_E$. But we have $\text{ang}(l, l_i) > \frac{\pi}{2} - \varepsilon = \arccos(\frac{1}{\sqrt{d}n})$ by Lemma 60. Thus, $l_i \notin PC_E$.

To see the second observation, let B be an orthogonal basis of $\mathbb{R}^d$ with $l \notin PC_B$. By Lemma 58, there exists $x \in PC_B$ with $\text{ang}(l, x) = \frac{\pi}{2}$. Then by Lemma 59, there exists $y \in PC_B$ such that $\text{ang}(x, y) < \varepsilon$ and the $\frac{\varepsilon}{3\sqrt{d}}$-ball at y is contained in PC_B. Since $\text{ang}(l, x) = \frac{\pi}{2}$ and $\text{ang}(x, y) < \varepsilon$, we have $\text{ang}(l, y) > \frac{\pi}{2} - \varepsilon$. Therefore, according to the condition (2) in Lemma 60, there must exist some l_i such that $\text{ang}(l_i, y) < \xi$. Recall that $\xi = \frac{\varepsilon}{3\sqrt{d}}$, so l_i is in the $\frac{\varepsilon}{3\sqrt{d}}$-ball at y and hence in PC_B. These two observations will be used later to verify that Ψ satisfies the desired property.

Now we continue to discuss the construction of Ψ. We have computed m points $l_1, \ldots, l_m \in \mathbb{P}^{d-1}$ for a specific pair (a, a'). We do the same thing for all pairs (a, a') of points in S with $\text{cl}(a) \neq \text{cl}(a')$ and $a \succ_E a'$. After this, we obtain $M = O(n^2 m) = O(n^d)$ points in $\mathbb{P}^{d-1}$ (with an abuse of notation, we denote them by $l_1, \ldots, l_M$). The set Ψ we construct consists of $2M$ points $b_1, \ldots, b_M, b'_1, \ldots, b'_M \in \mathbb{R}^d$ where b_i, b'_i correspond to l_i for $i \in \{1, \ldots, M\}$. We set the coordinates of each b_i in $\mathbb{R}^d$ to be $(-i, \ldots, -i, n+i)$. Then we choose location for each b'_i in $\mathbb{R}^d$ such that $\|b'_i - b_i\|_2 < 0.1$ and $\overline{b'_i - b_i} = l_i$ (there are infinitely many choices, we arbitrarily pick one of them). Finally, we need to define the coloring of the points in Ψ, i.e., $\text{cl}'(b)$ for all $b \in \Psi$. We arbitrarily color the points in Ψ under the only restriction that b_i and b'_i must have different colors, i.e., $\text{cl}'(b_i) \neq \text{cl}'(b'_i)$, for all $i \in \{1, \ldots, M\}$. It suffices to verify the property that $A \subseteq S$ contains inter-color dominances with respect to E iff $A \cup \Psi \subseteq S'$ contains inter-color dominances with respect to any orthogonal basis.

To see the "if" part, let $A \subseteq S$ be a subset such that $A \cup \Psi \subseteq S'$ contains inter-color dominances with respect to any orthogonal basis. Since $S \subset [1, n]^d$ (as S is regular) and $l_i \notin PC_E$ for all $i \in \{1, \ldots, M\}$ (as observed above), the points in Ψ do not dominate each other and do not form dominance with any points in S, with respect to E. But by assumption, $A \cup \Psi$ contains inter-color dominances with respect to E. So the inter-color dominances must be formed by the points in A, i.e.,

A contains inter-color dominances with respect to E.

To see the "only if" part, let $A \subseteq S$ be a subset containing inter-color dominances with respect to E. Suppose $a, a' \in A$ are two points such that $\mathrm{cl}(a) \neq \mathrm{cl}(a')$ and $a \succ_E a'$. Consider an orthogonal basis B of $\mathbb{R}^d$, and we must show that $A \cup \Psi$ contains inter-color dominances with respect to B. If $a \succ_B a'$ or $a' \succ_B a$, then we are done. Otherwise, recall that we have m points in $\{l_1, \dots, l_M\}$ which are chosen for the pair (a, a') (assume they are $l_1, \dots, l_m$ without loss of generality). By our observation above, one of these m points must be in PC_B, say $l_1 \in PC_B$. Then the two points $b_1, b_1' \in \Psi$ form an inter-color dominance with respect to B.

By the above construction, we obtain a colored stochastic dataset $\mathcal{S}' = (S \cup \Psi, \mathrm{cl}', \pi')$ in $\mathbb{R}^d$ satisfying $\Lambda^*_{\mathcal{S}'} = \Lambda_{\mathcal{S}}$. Clearly, this reduction can be done in polynomial time. Thus, the FBCSD problem is #P-hard for $d \geq 3$. In fact, with some efforts, one can make this result stronger by considering the FBCSD problem with respect to a balanced color pattern.

Theorem 61. *Let $\mathcal{P}' = (\Delta'_1, \Delta'_2, \dots)$ be a balanced color pattern. Then for any fixed d, there exists a balanced color pattern $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ such that the CSD problem in $\mathbb{R}^d$ with respect to $\mathcal{P}$ is polynomial-time reducible to the FBCSD problem in $\mathbb{R}^d$ with respect to $\mathcal{P}'$. In particular, the FBCSD problem in $\mathbb{R}^d$ with respect to $\mathcal{P}'$ is #P-hard for $d \geq 3$.*

Proof. To prove the result, we first determine some constants. Since $\mathcal{P}'$ is balanced, there is a constant $c_1 < 1$ such that $N - \max \Delta'_N \geq N^{c_1}$ for any sufficiently large N. Recall that our construction of the auxiliary point set Ψ satisfies $|\Psi| = 2M = O(n^d)$ where $n = |S|$. So we can find a constant c_2 such that $|S \cup \Psi| \leq c_2 n^d$. We construct the desired balanced color pattern $\mathcal{P} = (\Delta_1, \Delta_2, \dots)$ as follows. For an integer $p > 0$, in order to determine Δ_p, set $q = (c_2 p^d)^{2/c_1}$. We consider two cases, $|\Delta'_q| \geq c_2 p^d$ and $|\Delta'_q| < c_2 p^d$. In the case of $|\Delta'_q| \geq c_2 p^d$, we define $\Delta_p = \{1, \dots, 1\}$, i.e., a multi-set consisting of p 1's. In the case of $|\Delta'_q| < c_2 p^d$, we define $\Delta_p = \{\frac{p}{2}, \frac{p}{2}\}$ if p is even and $\Delta_p = \{\frac{p-1}{2}, \frac{p+1}{2}\}$ if p is odd. This completes the construction of $\mathcal{P}$.

We claim that the CSD problem in $\mathbb{R}^d$ with respect to $\mathcal{P}$ is polynomial-time reducible to the FBCSD problem in the same dimension with respect to $\mathcal{P}'$. Let $\mathcal{S} = (S, \mathrm{cl}, \pi)$ be a colored stochastic dataset in $\mathbb{R}^d$ such that $\langle \mathcal{S} \rangle$ is an instance of the CSD problem with respect to $\mathcal{P}$. Suppose $|S| = n$ and set $N = (c_2 n^d)^{2/c_1}$. We want

to construct another colored stochastic dataset $\mathcal{S}' = (S', \mathrm{cl}', \pi')$ in $\mathbb{R}^d$ with $|S'| = N$ such that $\Lambda^*_{\mathcal{S}'} = \Lambda_{\mathcal{S}}$ and $\langle \mathcal{S}' \rangle$ is an instance of the FBCSD problem with respect to $\mathcal{P}'$. As before, we first construct the auxiliary point set $\Psi = \{b_1, \ldots, b_M, b'_1, \ldots, b'_M\}$. By our assumption, we have $|S \cup \Psi| = n + 2M \leq c_2 n^d < N$. In order to have $|S'| = N$, we then arbitrarily choose a set D of $N - (n + 2M)$ dummy points in $\mathbb{R}^d$ (these points can be chosen arbitrarily as we will assign them existence probabilities 0 later) and set $S' = S \cup \Psi \cup D$. The existence probabilities of the points in S' are defined as

$$\pi'(a) = \begin{cases} \pi(a) & \text{if } a \in S, \\ 1 & \text{if } a \in \Psi, \\ 0 & \text{if } a \in D. \end{cases}$$

It suffices to define the coloring cl' of S'. Since we need $\langle \mathcal{S}' \rangle$ to be an instance of the FBCSD problem with respect to $\mathcal{P}'$, cl' must induce the partition Δ'_N. Besides, it should be guaranteed that $\mathrm{cl}'(a) = \mathrm{cl}(a)$ for all $a \in S$ and $\mathrm{cl}'(b_i) \neq \mathrm{cl}'(b'_i)$ for $i \in \{1, \ldots, M\}$ (as observed previously, $\Lambda^*_{\mathcal{S}'} = \Lambda_{\mathcal{S}}$ as long as we have this).

We consider two cases, $|\Delta'_N| \geq c_2 n^d$ and $|\Delta'_N| < c_2 n^d$. In the case of $|\Delta'_N| \geq c_2 n^d$, we have $\Delta_n = \{1, \ldots, 1\}$ by definition and therefore all the points in S have distinct colors (under the coloring cl). Note that $|S \cup \Psi| \leq c_2 n^d \leq |\Delta'_N|$. As such, one can easily find a coloring cl' inducing Δ'_N which assigns distinct colors to the points in $S \cup \Psi$ and satisfies $\mathrm{cl}'(a) = \mathrm{cl}(a)$ for all $a \in S$ (note that the coloring on D is "free", so we can easily make cl' induces Δ'_N). This cl' completes our reduction. In the case of $|\Delta'_N| < c_2 n^d$, we have that $\Delta_n = \{\frac{n}{2}, \frac{n}{2}\}$ if n is even and $\Delta_n = \{\frac{n-1}{2}, \frac{n+1}{2}\}$ if n is odd. Without loss of generality, we may assume that $\mathrm{cl}(S) = \{1, 2\}$. Suppose $\Delta'_N = \{r_1, \ldots, r_m\}$ where $m < c_2 n^d$ and $r_1 \geq \cdots \geq r_m$. We claim that $r_1 \geq r_2 \geq c_2 n^d$. Indeed, if $r_2 < c_2 n^d$, then $\sum_{i=2}^{m} r_i < m c_2 n^d < (c_2 n^d)^2$ and hence $N \leq (N - r_1)^{1/c_1} < (c_2 n^d)^{2/c_1}$, contradicting the fact that $N = (c_2 n^d)^{2/c_1}$. With this observation, we try to construct cl' with $\mathrm{cl}'(S') = \{1, \ldots, m\}$ such that cl' assigns color i to exactly r_i points in S'. We define $\mathrm{cl}'(a) = \mathrm{cl}(a)$ for all $a \in S$, $\mathrm{cl}'(b_i) = 1$ and $\mathrm{cl}'(b'_i) = 2$ for $i \in \{1, \ldots, M\}$. Note that by doing this we do not "exhaust" the colors 1 and 2, because $r_1 \geq r_2 \geq c_2 n^d \geq |S \cup \Psi|$. So we can carefully determine $\mathrm{cl}'(a)$ for all $a \in D$ such that exactly r_i points in S' have color i. By the defined cl', we completes our reduction and the proof. $\square$

4.3.2 Reduction to the CSD problem for $d = 2$

In this section, we study the FBCSD problem for $d = 2$ and show that an instance of the FBCSD problem in $\mathbb{R}^2$ can be reduced to $O(n^2)$ instances of the CSD problem in $\mathbb{R}^2$. By combining this reduction with our algorithm given in Section 4.2.1, we directly obtain an $O(n^4 \log^2 n)$-time algorithm for the FBCSD problem in $\mathbb{R}^2$. For simplicity of exposition, we assume that S is in general position in $\mathbb{R}^2$, i.e., no three points are collinear.

We try to compute $\Gamma^*_{\mathcal{S}}$. When computing $\Gamma^*_{\mathcal{S}}$, we need to consider the realizations of $\mathcal{S}$ which contain no inter-color dominances with respect to some orthogonal basis of $\mathbb{R}^2$ (these realizations are said to be *good*). We first establish a criterion for testing whether a realization is good. Recall that for a nonzero point $x \in \mathbb{R}^d$, the notation $\overline{x}$ denotes the image of x in $\mathbb{P}^{d-1}$ under the quotient map $\mathbb{R}^d \backslash \{\mathbf{0}\} \to \mathbb{P}^{d-1}$. For a subset $A \subseteq S$, we define $L_A = \{\overline{a_i - a_j} : a_i, a_j \in A \text{ and } \mathrm{cl}(a_i) \neq \mathrm{cl}(a_j)\} \subset \mathbb{P}^1$. For two points $l, l' \in \mathbb{P}^1$, we denote by $\theta(l, l')$ the angle between l and l' whose counterclockwise boundary is l and clockwise boundary is l' (when talking about angle we regard l and l' as lines in $\mathbb{R}^2$ through the origin). Then we have the following observation.

Lemma 62. *A realization R of $\mathcal{S}$ is good iff $L_R = \emptyset$ or there exists a unique $l \in L_R$ such that $\theta(l, l') > \frac{\pi}{2}$ for any $l' \in L_R$ not equal to l.*

Proof. We first consider the "if" part. If $L_R = \emptyset$, then R is monochromatic and hence is good. If there exists $l \in L_R$ such that $\theta(l, l') > \frac{\pi}{2}$ for any $l' \in L_R$ not equal to l, one can slightly rotate l clockwise to obtain $l_0 \in \mathbb{P}^1$ such that $\theta(l_0, l') > \frac{\pi}{2}$ for any $l' \in L_R$. Suppose the homogeneous coordinates of l_0 is $[\alpha : \beta]$ with $\alpha^2 + \beta^2 = 1$. Take the orthogonal basis $B = (\mathbf{b}_1, \mathbf{b}_2)$ of $\mathbb{R}^2$ with $\mathbf{b}_1 = (\alpha, \beta)$ and $\mathbf{b}_2 = (\beta, -\alpha)$. Since $\theta(l_0, l') > \frac{\pi}{2}$ for any $l' \in L_R$, we know that R contains no inter-color dominances with respect to B.

To see the "only if" part, let R be a good realization of $\mathcal{S}$. Suppose R contains no inter-color dominances with respect to some orthogonal basis $B = (\mathbf{b}_1, \mathbf{b}_2)$ of $\mathbb{R}^2$ (assume $\mathbf{b}_2$ is in the clockwise direction of $\mathbf{b}_1$ with angle $\frac{\pi}{2}$). Let $b \in \mathbb{P}^1$ be the point corresponding to $\mathbf{b}_1$ (i.e., b is the image of $\mathbf{b}_1$ under the quotient map $S^1 \to \mathbb{P}^1$). If $L_R = \emptyset$, we are done. So assume $L_R \neq \emptyset$. Define $l \in L_R$ as the point which minimizes $\theta(l, b)$. We claim that $\theta(l, l') > \frac{\pi}{2}$ for any $l' \in L_R$ not equal to l.

Let $l' \in L_R$ be a point not equal to l. If $\theta(l, l') \leq \frac{\pi}{2}$, then either $\theta(l', b) < \theta(l, b)$ or $l' \in PC_B$ (recall that PC_B is the projective cone of B defined in Section 4.3.1). The former contradicts the definition of l while the latter contradicts the fact that R contains no inter-color dominances with respect to B. □

Note that $L_R = \emptyset$ iff R is monochromatic. Based on the above lemma, we now define a notion called *witness pair* as follows. Let R be a good but not monochromatic realization of $\mathcal{S}$. Then by Lemma 62, there exists a unique $l \in L_R$ such that $\theta(l, l') > \frac{\pi}{2}$ for any $l' \in L_R$ not equal to l. According to the definition of L_R, we must have $l = \overline{a_i - a_j}$ for some $a_i, a_j \in R$ with $\text{cl}(a_i) \neq \text{cl}(a_j)$. Note that the choice of a_i, a_j is not necessarily unique (though l is unique). Let Y be the set of all pairs (a_i, a_j) with $a_i, a_j \in R$ satisfying $\text{cl}(a_i) \neq \text{cl}(a_j)$ and $l = \overline{a_i - a_j}$. We claim that there exists a unique pair $(a_{i^*}, a_{j^*}) \in Y$ such that for any $(a_i, a_j) \in Y$ we have $j^* \geq j$. The existence is obvious, so it suffices to show the uniqueness. Indeed, if (a_i, a_j) and $(a_{i'}, a_j)$ are two pairs in Y, then the points $a_i, a_{i'}, a_j$ must be collinear in $\mathbb{R}^2$. However, because of the general position assumption for S (and hence for R), we must have $i = i'$. It follows that for any $j \in \{1, \ldots, n\}$ there is at most one pair $(a_i, a_j) \in Y$, which further implies the uniqueness of (a_{i^*}, a_{j^*}). We define the pair (a_{i^*}, a_{j^*}) as the *witness pair* of R, denoted by $\text{wit}(R)$. See Figure 4.12 for an example. Now it is clear that

$$\Gamma_S^* = Pr_{\text{mono}} + \sum_{i=1}^{n} \sum_{j=1}^{n} Pr_{i,j},$$

where Pr_{mono} is the probability that a realization R of $\mathcal{S}$ is monochromatic, and $Pr_{i,j}$ is the probability that R is good (but not monochromatic) with $\text{wit}(R) = (a_i, a_j)$.

It is easy to compute Pr_{mono} in linear time. The problem remaining is how to compute $Pr_{i,j}$ for all $i, j \in \{1, \ldots, n\}$. Fix a pair (i^*, j^*). Obviously, if $\text{cl}(a_{i^*}) = \text{cl}(a_{j^*})$, we immediately have $Pr_{i^*,j^*} = 0$. So suppose $\text{cl}(a_{i^*}) \neq \text{cl}(a_{j^*})$. We try to reduce the the task of computing Pr_{i^*,j^*} to an instance of the CSD problem in $\mathbb{R}^2$. Let $\mathbf{b}_1 = (a_{i^*} - a_{j^*})/\|a_{i^*} - a_{j^*}\|_2$ be a unit vector of $\mathbb{R}^2$, and $\mathbf{b}_2$ be another unit vector obtained by rotating $\mathbf{b}_1$ clockwise with angle $\frac{\pi}{2}$. Clearly, $B = (\mathbf{b}_1, \mathbf{b}_2)$ is an orthogonal basis of $\mathbb{R}^2$. We define n points $a'_1, \ldots, a'_n \in \mathbb{R}^2$ as follows. Let δ be a small enough real number such that for any $i, j \in \{1, \ldots, n\}$ we have $|\langle \mathbf{b}_2, a_i \rangle - \langle \mathbf{b}_2, a_j \rangle| > \delta$

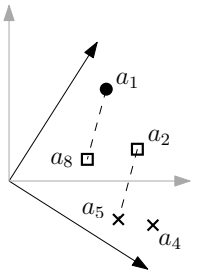


Figure 4.12: An example of witness pair. $l = \overline{a_1 - a_8} = \overline{a_2 - a_5}$. $\text{wit}(R) = (a_1, a_8)$.

unless $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$. Consider a specific index $p \in \{1, \ldots, n\}$. If $p \leq j^*$ and there exists $q \leq j^*$ satisfying $\text{cl}(a_p) \neq \text{cl}(a_q)$, $a_p \succ_B a_q$, $\langle \mathbf{b}_2, a_p \rangle = \langle \mathbf{b}_2, a_q \rangle$, then we set the coordinates of a'_p in $\mathbb{R}^2$ to be $(\langle \mathbf{b}_2, a_p \rangle - \delta, \langle \mathbf{b}_1, a_p \rangle)$. Otherwise, we set the coordinates of a'_p to be $(\langle \mathbf{b}_2, a_p \rangle, \langle \mathbf{b}_1, a_p \rangle)$. Based on this, we can construct a colored stochastic dataset $\mathcal{S}' = (S', \text{cl}', \pi')$ in $\mathbb{R}^2$ by defining $S' = \{a'_1, \ldots, a'_n\}$, $\text{cl}'(a'_i) = \text{cl}(a_i)$ for all $i \in \{1, \ldots, n\}$, and $\pi'(a'_{i^*}) = \pi'(a'_{j^*}) = 1$, $\pi'(a'_i) = \pi(a_i)$ for all $i \in \{1, \ldots, n\} \backslash \{i^*, j^*\}$. We observe the following equation, which allows us to compute Pr_{i^*,j^*} by solving the instance $\langle \mathcal{S}' \rangle$ of the CSD problem in $\mathbb{R}^2$.

Lemma 63. $Pr_{i^*,j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Gamma_{\mathcal{S}'}$.

Proof. First, we observe (Observation 1, hereafter) that a realization R of $\mathcal{S}$ is good with $\text{wit}(R) = (a_{i^*}, a_{j^*})$ iff (1) $a_{i^*}, a_{j^*} \in R$ and (2) for any $a_i, a_j \in R$ with $\text{cl}(a_i) \neq \text{cl}(a_j)$ and $a_i \succ_B a_j$, we have $\max(i, j) \leq j^*$ and $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$. To see the "if" part, assume R satisfies the conditions (1) and (2). Since $a_{i^*}, a_{j^*} \in R$, we know that $\overline{a_{i^*} - a_{j^*}} \in L_R$. Set $l = \overline{a_{i^*} - a_{j^*}}$. The condition (2) guarantees that $\theta(l, l') > \frac{\pi}{2}$ for any $l' \in L_R$ not equal to l. Thus, by Lemma 62, R is good (but not monochromatic since $a_{i^*}, a_{j^*} \in R$). Furthermore, it is easy to see that the conditions (1) and (2) also guarantee $\text{wit}(R) = (a_{i^*}, a_{j^*})$. To see the "only if" part, assume R is good with $\text{wit}(R) = (a_{i^*}, a_{j^*})$. By the definition of witness pair, we immediately have $a_{i^*}, a_{j^*} \in R$. Again, set $l = \overline{a_{i^*} - a_{j^*}}$. Let $a_i, a_j \in R$ be two points such that $\text{cl}(a_i) \neq \text{cl}(a_j)$ and $a_i \succ_B a_j$. By the definition of B, we have $\theta(l, l') \leq \frac{\pi}{2}$ for $l' = \overline{a_i - a_j}$. According to Lemma 62, it implies that $l = l'$, i.e., $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$. Besides, we must have $\max(i, j) \leq j^*$, otherwise (a_{i^*}, a_{j^*}) is not the witness pair of R.

Second, we observe (Observation 2, hereafter) that our construction of $\mathcal{S}'$ satisfies

the following property. Let $i, j \in \{1, \ldots, n\}$ be any indices such that $\mathrm{cl}(a_i) \neq \mathrm{cl}(a_j)$, or equivalently, $\mathrm{cl}'(a'_i) \neq \mathrm{cl}'(a'_j)$. Then we have $a'_i \succ_E a'_j$ iff (1) $a_i \succ_B a_j$ and (2) $\langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle$ or $\max(i, j) > j^*$. To see the "if" part, assume $a_i \succ_B a_j$. In this case, we have $y(a'_i) = \langle \mathbf{b}_1, a_i \rangle \geq \langle \mathbf{b}_1, a_j \rangle = y(a'_j)$. If $\langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle$, then $x(a'_i) \geq \langle \mathbf{b}_2 - \delta, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle \geq x(a'_j)$ so that $a'_i \succ_E a'_j$. If $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ and $\max(i, j) > j^*$, we also have $x(a'_i) = \langle \mathbf{b}_2, a_i \rangle > \langle \mathbf{b}_2, a_j \rangle = x(a'_j)$ (recall the general position assumption) so that $a'_i \succ_E a'_j$. To see the "only if" part, first assume $a_i \not\succ_B a_j$. In this case, we have either $y(a'_i) < y(a'_j)$ or $x(a'_i) < x(a'_j)$, which implies $a'_i \not\succ_E a'_j$. Now assume $a_i \succ_B a_j$, $\langle \mathbf{b}_2, a_i \rangle \leq \langle \mathbf{b}_2, a_j \rangle$, and $\max(i, j) \leq j^*$. Because $a_i \succ_B a_j$, it must be the case that $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ and $\langle \mathbf{b}_1, a_i \rangle > \langle \mathbf{b}_1, a_j \rangle$. By our construction, we have $x(a'_i) = \langle \mathbf{b}_2, a_i \rangle - \delta$. But $x(a'_j) = \langle \mathbf{b}_2, a_j \rangle$ (recall the general position assumption). Thus, $a'_i \not\succ_E a'_j$.

With the above two observations, we prove the equation $Pr_{i^*,j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Gamma_{\mathcal{S}'}$. Define a natural one-to-one correspondence $\mu : S \to S'$ as $\mu(a_i) = a'_i$. First, it is clear that for any subset $A \subseteq S$ including a_{i^*}, a_{j^*}, the probability that A occurs as a realization of $\mathcal{S}$ is equal to the product $\pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Pr[\mu(A)]$, where $\Pr[\mu(A)]$ the probability that $\mu(A)$ occurs as a realization of $\mathcal{S}'$. Let R be a realization of $\mathcal{S}$. We claim that R is good with $\mathrm{wit}(R) = (a_{i^*}, a_{j^*})$ iff $a'_{i^*}, a'_{j^*} \in \mu(R)$ and $\mu(R)$ contains no inter-color dominances (with respect to E).

To see the "if" part, assume $a'_{i^*}, a'_{j^*} \in \mu(R)$ and $\mu(R)$ contains no inter-color dominances with respect to E. Then $a_{i^*}, a_{j^*} \in R$. Let $a_i, a_j \in R$ be two points such that $\mathrm{cl}(a_i) \neq \mathrm{cl}(a_j)$ and $a_i \succ_B a_j$. Since $\mu(R)$ contains no inter-color dominances with respect to E, Observation 2 above implies that $\max(i, j) \leq j^*$ and $\langle \mathbf{b}_2, a_i \rangle \leq \langle \mathbf{b}_2, a_j \rangle$ (the latter further implies $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$ since $a_i \succ_B a_j$). Thus, by Observation 1 above, R is good with $\mathrm{wit}(R) = (a_{i^*}, a_{j^*})$.

To see the "only if" part, assume R is good with $\mathrm{wit}(R) = (a_{i^*}, a_{j^*})$. Then Observation 1 implies $a_{i^*}, a_{j^*} \in R$ and hence $a'_{i^*}, a'_{j^*} \in \mu(R)$. Let $a_i, a_j \in R$ be two points such that $\mathrm{cl}(a_i) \neq \mathrm{cl}(a_j)$. If $a_i \not\succ_B a_j$, then by Observation 2 we have $a'_i \not\succ_E a'_j$. If $a_i \succ_B a_j$, then Observation 1 implies that $\max(i, j) \leq j^*$ and $\langle \mathbf{b}_2, a_i \rangle = \langle \mathbf{b}_2, a_j \rangle$. By using Observation 2, we also have $a'_i \not\succ_E a'_j$. Therefore, $\mu(R)$ contains no inter-color dominances with respect to E. This argument shows that μ induces a one-to-one correspondence between the good realizations of $\mathcal{S}$ and the realizations of $\mathcal{S}'$ which include a'_{i^*}, a'_{j^*} and contain no inter-color dominances (with

respect to E). Note that $\pi'(a'_{i^*}) = \pi'(a'_{j^*}) = 1$, hence a realization of $\mathcal{S}'$ for sure includes a'_{i^*}, a'_{j^*}. As a result, we have $Pr_{i^*,j^*} = \pi(a_{i^*}) \cdot \pi(a_{j^*}) \cdot \Gamma_{\mathcal{S}'}$. □

In this way, an instance of the FBCSD problem in $\mathbb{R}^2$ is reduced to $O(n^2)$ instances of the CSD problem in $\mathbb{R}^2$. By plugging in our $O(n^2 \log^2 n)$ algorithm for solving the CSD problem in $\mathbb{R}^2$, we have the following result.

Theorem 64. *The FBCSD problem in $\mathbb{R}^2$ can be solved in $O(n^4 \log^2 n)$ time.*

Chapter 5

Conclusion and future work

5.1 Conclusion

In this thesis, we investigated three classes of geometric problems on stochastic datasets that are equipped with existential uncertainty. The first class of problems considers the linear separability of a bichromatic stochastic dataset, the second class of problems considers the expected measures of a stochastic convex hull, and the third class of problems considers the dominance relation in a colored stochastic dataset.

For the stochastic separability, we studied the separable-probability (SP) problem and the expected separation-margin (ESM) problem, which are defined in Section 1.1. We designed efficient algorithms for computing the SP and ESM of a bichromatic stochastic dataset, and also provided hardness results for these problems.

For the stochastic convex hull, we studied the problem of computing the expected diameter, width, and combinatorial complexity of a SCH of a stochastic dataset. We gave efficient approximation algorithms for the expected diameter and width problems, and an exact algorithm for the expected complexity problem. Also, we showed that exactly computing the expected diameter is #P-hard when the dimension is not fixed.

For the stochastic dominance, we studied the colored stochastic dominance

(CSD) problem and the free-basis colored stochastic dominance (FBCSD) problem for a colored stochastic dataset. We established efficient algorithms for both problems when $d = 2$ and showed #P-hardness for both problems when $d \geq 3$. Also, we gave an FPRAS for the CSD problem in any dimension.

In sum, our results demonstrated that geometric problems on uncertain (i.e., stochastic) datasets are significantly more difficult than their counterparts on conventional datasets: many problems that are efficiently solvable on conventional datasets require much more time to be solved or even become #P-hard on uncertaint datasets.

5.2 Future work

In this section, we list some directions for future study on geometric computing in stochastic settings. For the stochastic separability-related problems, one direction for future study is to propose efficient approximation algorithms. As we have seen in Section 2, the running times of our algorithms are exponential in d, and it might be difficult to further improve our algorithms. However, if we only want an approximation of the SP or ESM, it might be possible to design more efficient algorithms. For example, one may seek a constant-approximation algorithm or even a $(1 + \varepsilon)$-approximation algorithm for computing the SP or ESM; whether there exist such algorithms with time complexity polynomial in n, N, and d is still an open question. Another direction is to study other problems related to the separability of a bichromatic stochastic dataset. For example, one can study the problem of finding the *most likely* maximum-margin separator of a given bichromatic stochastic dataset, i.e., the hyperplane with the maximum probability of being the maximum-margin separator of a realization.

For the stochastic convex hull-related problems, one direction for future study is to propose fully polynomial-time approximation schemes (FPTAS), i.e., deterministic $(1 + \varepsilon)$-approximation algorithms whose time complexity is polynomial in both n and $1/\varepsilon$, for computing the expected diameter and width of a SCH. Another direction is to prove hardness results for computing the expected diameter and width. We have proved in Section 3.2.4 that computing the expected diameter exactly when d is not fixed is #P-hard. However, when d is fixed, it is not clear

whether the problem is NP-hard/#P-hard or polynomial-time solvable. Also, no hardness result for the expected-width problem is currently known.

For the stochastic dominance-related problems, one open question is whether the hardness result presented in Section 4.2.2.3 can be further extended to the case of $d \geq 3$. Another direction for future study is to further improve the time complexity of the algorithms in Section 4.2.1 and 4.3.2.

Besides the aforementioned problems, one can also investigate other kinds of geometric problems on stochastic datasets. Here we list some potential problems for future study.

Stochastic range-counting problems. Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset in $\mathbb{R}^d$. For a range $Q \subseteq \mathbb{R}^d$, define a random variable $n_Q = |R \cap Q|$ where R is a realization of $\mathcal{S}$. The stochastic range-counting problems aim to preprocess $\mathcal{S}$ into some data structure such that given a query range Q, the information about n_Q can be computed efficiently. For example, one may ask for the expectation of n_Q, the probability that n_Q equals to (or greater/smaller than) a specified constant k, etc.

Stochastic connecting-distance problems. For a set S of points in $\mathbb{R}^d$, the *connecting distance* of S is the smallest real number δ such that for any $a, a' \in S$, there is a sequence $b_1, \ldots, b_r$ of points in S satisfying $\text{dist}(a, b_1) \leq \delta$, $\text{dist}(b_r, a') \leq \delta$, and $\text{dist}(b_i, b_{i+1}) \leq \delta$ for all $i \in \{1, \ldots, d-1\}$. Let $\mathcal{S} = (S, \pi)$ be a stochastic dataset in $\mathbb{R}^d$. The stochastic connecting-distance problems involve computing the information about the connecting distance of a realization of $\mathcal{S}$ (which is a random variable). For example, one may ask for the expected connecting distance, the probability that the connecting distance equals to (or greater/smaller than) a specified constant k, etc.

References

[1] Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Cheong Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, 2000.

[2] Franco P Preparata and Michael I Shamos. *Computational Geometry: An Introduction*. Springer Science & Business Media, 2012.

[3] Raimund Seidel. Convex hull computations. *Handbook of Discrete and Computational Geometry (edited by Jacob Goodman and Joseph O'Rourke)*, pages 495–512, 2004.

[4] Pankaj K Agarwal, Boris Aronov, Sariel Har-Peled, Jeff M Phillips, Ke Yi, and Wuzhou Zhang. Nearest-neighbor searching under uncertainty II. *ACM Transactions on Algorithms*, 13(1):3, 2016.

[5] Subhash Suri and Kevin Verbeek. On the most likely voronoi diagram and nearest neighbor searching. *International Journal of Computational Geometry & Applications*, 26:151–166, 2016.

[6] Jie Xue and Yuan Li. Stochastic closest-pair problem and most-likely nearest-neighbor search in tree spaces. In *Proceedings of the 15th International Symposium on Algorithms and Data Structures*, pages 569–580. Springer, 2017.

[7] Pankaj K Agarwal, Sariel Har-Peled, Subhash Suri, Hakan Yıldız, and Wuzhou Zhang. Convex hulls under uncertainty. In *Proceedings of the 22nd European Symposium on Algorithms*, pages 37–48. Springer, 2014.

[8] Lingxiao Huang and Jian Li. Approximating the expected values for combinatorial optimization problems over stochastic points. In *Proceedings of the 42nd*

International Colloquium on Automata, Languages, and Programming, pages 910–921. Springer, 2015.

[9] Chao Li, Chenglin Fan, Jun Luo, Farong Zhong, and Binhai Zhu. Expected computations on color spanning sets. *Journal of Combinatorial Optimization*, 29(3):589–604, 2015.

[10] Maarten Löffler and Marc van Kreveld. Largest and smallest convex hulls for imprecise points. *Algorithmica*, 56(2):235, 2010.

[11] Subhash Suri, Kevin Verbeek, and Hakan Yıldız. On the most likely convex hull of uncertain points. In *Proceedings of the 21st European Symposium on Algorithms*, pages 791–802. Springer, 2013.

[12] Jie Xue, Yuan Li, and Ravi Janardan. On the expected diameter, width, and complexity of a stochastic convex hull. *Computational Geometry: Theory and Applications*, 82:16–31, 2019.

[13] Pegah Kamousi, Timothy M Chan, and Subhash Suri. Stochastic minimum spanning trees in euclidean spaces. In *Proceedings of the 27th Annual Symposium on Computational geometry*, pages 65–74. ACM, 2011.

[14] Pegah Kamousi, Timothy M Chan, and Subhash Suri. Closest pair and the post office problem for stochastic points. *Computational Geometry: Theory and Applications*, 47(2):214–223, 2014.

[15] Pankaj K Agarwal, Siu-Wing Cheng, and Ke Yi. Range searching on uncertain data. *ACM Transactions on Algorithms*, 8(4):43, 2012.

[16] Pankaj K Agarwal, Nirman Kumar, Stavros Sintos, and Subhash Suri. Range-max queries on uncertain data. *Journal of Computer and System Sciences*, 94:118–134, 2018.

[17] Lingxiao Huang and Jian Li. Stochastic k-center and j-flat-center problems. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 110–129. SIAM, 2017.

[18] Nirman Kumar, Benjamin Raichel, Subhash Suri, and Kevin Verbeek. Most likely voronoi diagrams in higher dimensions. In *Proceedings of the 36th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2016)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2016.

[19] Yuan Li, Jie Xue, Akash Agrawal, and Ravi Janardan. On the arrangement of stochastic lines in $\mathbb{R}^2$. *Journal of Discrete Algorithms*, 44:1–20, 2017.

[20] Mark de Berg, Ali D Mehrabi, and Farnaz Sheikhi. Separability of imprecise points. In *Proceedings of the 14th Scandinavian Workshop on Algorithm Theory*, pages 146–157. Springer, 2014.

[21] Martin Fink, John Hershberger, Nirman Kumar, and Subhash Suri. Hyperplane separability and convexity of probabilistic point sets. *Journal of Computational Geometry*, 8(2):32–57, 2017.

[22] Jie Xue, Yuan Li, and Ravi Janardan. On the separability of stochastic geometric objects, with applications. In *Proceedings of the 32nd Annual Symposium on Computational Geometry*. ACM, 2016.

[23] Akash Agrawal, Yuan Li, Jie Xue, and Ravi Janardan. The most-likely skyline problem for stochastic points. In *Proceedings of the 29th Canadian Conference on Computational Geometry*, pages 78–83, 2017.

[24] Jie Xue and Yuan Li. Colored stochastic dominance problems. *arXiv preprint arXiv:1612.06954*, 2016.

[25] Allan Jørgensen, Maarten Löffler, and Jeff M Phillips. Geometric computations on indecisive points. In *Proceedings of the 12th Workshop on Algorithms and Data Structures*, pages 536–547. Springer, 2011.

[26] Harold N Gabow, Jon Louis Bentley, and Robert E Tarjan. Scaling and related techniques for geometry problems. In *Proceedings of the 16th Symposium on Theory of Computing*, pages 135–143. ACM, 1984.

[27] Hsiang-Tsung Kung, Fabrizio Luccio, and Franco P Preparata. On finding the maxima of a set of vectors. *Journal of the ACM*, 22(4):469–476, 1975.

[28] Christos H Papadimitriou and Mihalis Yannakakis. Multiobjective query optimization. In *Proceedings of the 20th SIGMOD Symposium on Principles of Database Systems*, pages 52–59. ACM, 2001.

[29] Peyman Afshani, Pankaj K Agarwal, Lars Arge, Kasper Green Larsen, and Jeff M Phillips. (Approximate) uncertain skylines. *Theory of Computing Systems*, 52(3):342–366, 2013.

[30] Jian Pei, Bin Jiang, Xuemin Lin, and Yidong Yuan. Probabilistic skylines on uncertain data. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 15–26. VLDB Endowment, 2007.

[31] Wenjie Zhang, Xuemin Lin, Ying Zhang, Muhammad Aamir Cheema, and Qing Zhang. Stochastic skylines. *ACM Transactions on Database Systems*, 37(2):14, 2012.

[32] Herbert Edelsbrunner and Leonidas J Guibas. Topologically sweeping an arrangement. In *Proceedings of the 18th Symposium on Theory of Computing*, pages 389–403. ACM, 1986.

[33] Gill Barequet and Sariel Har-Peled. Efficiently approximating the minimum-volume bounding box of a point set in three dimensions. *Journal of Algorithms*, 38(1):91–109, 2001.

[34] Sariel Har-Peled. *Geometric approximation algorithms*. Number 173. American Mathematical Society, 2011.

[35] Mingji Xia and Wenbo Zhao. #3-regular bipartite planar vertex cover is #P-complete. In *Proceedings of the 2nd International Conference on Theory and Applications of Models of Computation*, pages 356–364. Springer, 2006.

[36] Leslie G Valiant. Universality considerations in VLSI circuits. *IEEE Transactions on Computers*, 100(2):135–140, 1981.

[37] Neil Robertson, Daniel P Sanders, Paul Douglas Seymour, and Robin Thomas. Efficiently four-coloring planar graphs. In *Proceedings of the 28th Symposium on Theory of Computing*, pages 571–575. ACM, 1996.